Fall 2020

Estimation and Inference Under Model Uncertainty

Yizheng Wei

Estimation and Inference Under Model Uncertainty

by

Yizheng Wei

Bachelor of Science
Sichuan University, 2010

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2020

Accepted by:

Yanyuan Ma, Major Professor

Brian Habing, Committee Member

David Hitchcock, Committee Member

Xiaoyan Lin, Committee Member

Cherly Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I deeply appreciate Professor Yanyuan Ma, my advisor and mentor for my entire doctoral training. She constantly supported, guided and nurtured me through every stage of my research. Whenever I have questions, no matter big or small, Professor Ma always patiently answered them with her wisdom and deep understanding of statistics in a swift fashion. Even though we were not physically in the same place in the last four years, I never felt apart. There has been a few times that I feel frustrated, drifted away and not being productive, instead of giving up on me, Professor Ma shielded me with more flexibility and allowed me to have time to sort things out. Professor Ma also intrigued my interests in research by exposing me to various topics in statistics, giving me chances to collaborate with researchers from different institutions. When I was struggling with the feedback from the reviewers, Professor Ma showed me how to generalize these questions and respond to them from a higher prospective. When my code didn't work the way it should be, Professor Ma gave me useful tips to help me debug and improve working efficiency. When I write the draft of our manuscripts that initially might confused the readers, Professor Ma instructed me how to write manuscripts in an elegant and clear way. I considered myself fortunate enough to have Professor Ma as my advisor.

There are many other people helped me a lot during my long journey for Ph.D. I want to express my sincere gratitude to Dr.Tanya P. Garcia, Dr.Xinyu Zhang, Dr.Jinbo Chen, Dr.Samiran Sinha for their insightful advice and tireless help on my research. I also want to thank Dr.Xiaoyan Lin, Dr.David Hitchcock, Dr.Brian Habing for being my committee members and give me valuable comments on my dissertation

and presentation. I also would like to thank all the faculty members in department of statistics for teaching me various statistics courses and provide me abundant academic resources.

Moreover, I want to thank my friends, Shiwen, Ge, Jianxuan, Qianqian, Seungchul, Taeho, Xinchu, Xichen, Xiang, and many more in USC who made my life so colorful in Columbia. I will always cherish the memory with all of you.

Finally, I am grateful to my family: my wife Lu Wang, thank you for your love and support for the past eleven years. My father Mingbin Wei and my mother Min Zeng, thank you for your love, nurturing me the curiosity about the nature, the resilient of conquering various difficulties in my life, and all the happy memories. I want to present this dissertation as the best gift to you.

Abstract

Chapter 1 of this dissertation proposes a consistent and locally efficient estimator to estimate the model parameters for a logistic mixed effect model with random slopes. Our approach relaxes two typical assumptions: the random effects being normally distributed, and the covariates and random effects being independent of each other. Adhering to these assumptions is particularly difficult in health studies where in many cases we have limited resources to design experiments and gather data in long-term studies, while new findings from other fields might emerge, suggesting the violation of such assumptions. So it is crucial if we could have an estimator robust to such violations and then we could make better use of current data harvested using various valuable resources. Our method generalizes the framework presented in Garcia & Ma (2016) which also deals with a logistic mixed effect model but only considers a random intercept. A simulation study reveals that our proposed estimator remains consistent even when the independence and normality assumptions are violated. This contrasts from the traditional maximum likelihood estimator which is likely to be inconsistent when there is dependence between the covariates and random effects. Application of this work to a Huntington disease study reveals that disease diagnosis can be further improved using assessments of cognitive performance.

When a model of main research interest shares partial parameters with several other models, it is of benefit to incorporate the information contained in these other models to improve the estimation and prediction for the main model of interest. Various methods are possible to make use of the additional models as well as the additional observations related to these models. In Chapter 2, we propose an optimal strategy of

doing so in terms of prediction. We develop a fusion learning method that fuses the model averaging methodology with meta analysis and obtain the optimal weights. We also establish theory to support the method and show its desirable properties both when the main model is correct and when it is incorrect. Numerical experiments including simulation studies and data analysis are conducted to demonstrate the superior performance of our methods.

In Chapter 3, we propose a new pseudo-likelihood approach to fitting logistic regression models with two-phase data that has incomplete data structure. The existing methods included inverse probability weighted (IPW) methods, pseudo-likelihood (PL) methods, and maximum likelihood (ML) methods. MLEs either require that the complete phase I covariates be discrete with a small number of levels or of low dimension, or the continuous phase I covariates could be stratified properly. Therefore, they may not be able to make full use of the complete covariate information. In comparison, our method does not require to stratify the continuous phase I covariates, and is more resilient to the misclassified phase I covariates. And it could handle a relatively larger number of phase I covariates when the sample size is relatively small, in this case, MLEs may not have enough samples in certain strata to obtain a valid estimation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Consistent Estimator for Logistic Mixed Effect Models with Unknown Random Effect Structure

1.1 Introduction

A mixed effect logistic model is commonly used for analyzing clustered binary data arising in longitudinal studies of behavioral, social, health, and biomedical science. In the mixed effect logistic model, the logit of the success probability of the response is modeled as a linear function of fixed and random effect components. The observed data are $(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij})$, $j = 1, \cdots, m_i$ and $i = 1, \cdots, n$, where Y_{ij} is the binary response variable, $\mathbf{X}_{ij}$ is a p-vector that exerts a fixed effect and $\mathbf{Z}_{ij}$ is a q-dimensional random variable that has a random effect $\mathbf{R}_i \in \mathcal{R}^q$. Here, i and j denote the index for clusters and the subject within a cluster, respectively. The random effect is completely unobserved and we assume $m_i > q$ for identifiability for all i. This identifiability requirement will become self-evident in Section 1.2. The mixed effect logistic model is

$$\text{pr}(Y_{ij} = 1 | \mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{R}_i) = \frac{\exp(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{R}_i)}{1 + \exp(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{R}_i)}, \; j = 1, \cdots, m, \; i = 1, \cdots, n, \quad (1)$$

and the main objective is to consistently estimate the p-dimensional regression coefficient $\boldsymbol{\beta}$ in the presence of the unobserved random effect.

The standard maximum likelihood approach estimates $\boldsymbol{\beta}$ assumes that $\mathbf{R}_i$ has a parametric distribution (e.g., multivariate normal with zero mean and positive definite

variance-covariance matrix) and is independent of the covariates $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$. When the distribution for $\mathbf{R}_i$ is misspecified, however, the approach can yield biased parameter estimates and distorted type-I error rates (Heagerty & Kurland 2001, Agresti et al. 2004, Litiére et al. 2007, 2008). The misspecification may occur in terms of misspecifying the shape of the distribution, incorrectly assuming independence between the covariates and the random effect, or incorrectly assuming independence between the cluster size and the random effect. A good review on the potential bias due to misspecification of the distribution of $\mathbf{R}_i$ can be found in Neuhaus et al. (2011).

More flexible models for the distribution of $\mathbf{R}_i$ have been considered to circumvent the misspecification bias, but under limited settings. For linear mixed models, Zhang & Davidian (2001) proposed a smooth semi-nonparametric probability density for random effect and Zhang et al. (2008) proposed a negatively skewed random effect density. However, extending either method to generalized linear models is non-trivial, and imposing smoothness constraints or a skewness condition introduces computational complexities that we can actually avoid.

In this paper, we propose estimating parameters in the mixed effect logistic model without imposing any distributional assumptions on the random effect. Taking a semiparametric approach, we treat the distribution of $\mathbf{R}_i$ as a nuisance parameter and demonstrate that consistent estimates of $\boldsymbol{\beta}$ are obtained regardless of how the distribution of $\mathbf{R}_i$ is specified. We thus avoid unnecessary assumptions, such as a particular distributional shape for $\mathbf{R}_i$ (Zhang & Davidian 2001, Zhang et al. 2008) and the independence between covariates and the random effect $\mathbf{R}_i$. Our method generalizes the framework presented by Garcia & Ma (2016) which also deals with a logistic mixed effect model but only considered a random intercept. The presence of the random slope terms in our model means that their method no longer applies. Extending the result from random intercept to random slope is not as straightforward as it seems.

The rest of this paper is organized as follows. In Section 1.2, we develop semiparametric efficient estimator for $\boldsymbol{\beta}$. We demonstrate that the proposed estimator is consistent regardless of the assumed model for the distribution of $\mathbf{R}_i$, and the estimator achieves the asymptotic efficiency when the distribution for $\mathbf{R}_i$ is correctly modeled. In Section 1.3, we demonstrate through extensive simulation studies that the proposed estimator is robust to different distributional assumptions of $\mathbf{R}_i$, including different distributional shapes and dependence structures with covariates. The robustness property of the new estimator contrasts to the large biases of the maximum likelihood estimator when the distribution of $\mathbf{R}_i$ is misspecified. In Section 1.4, we apply our method to analyze a dataset from a study of Huntington disease and discover that the maximum likelihood estimator may result in misleading results about the importance of cognitive measures in relationship to diagnosis of Huntington disease. In contrast, our method detects one more cognitive measure crucial in determining the diagnostic result of Huntington disease. The paper ends with a brief discussion in Section 1.5. All technical details are given in an Appendix.

1.2 Main Results

1.2.1 Notation and assumptions

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im})^{\mathrm{T}}$ denote a m-dimensional vector, $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im})$ denote a $p \times m$ matrix, and $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{im})$ denote a $q \times m$ matrix. Without loss of generality, assume that the first q columns of $\mathbf{Z}_i$ form an invertible matrix.

Let f to denote various densities described by the subindices. The likelihood for the ith cluster formed by the model in Equation (1) is

$$
\begin{aligned}
& f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) = \int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i \mid \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i) d\mu(\mathbf{r}_i) \\
= & \int \prod_{j=1}^{m} \exp[y_{ij}(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{r}_i^{\mathrm{T}}\mathbf{z}_{ij}) - \log\{1 + \exp(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{r}_i^{\mathrm{T}}\mathbf{z}_{ij}\}] f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i) d\mu(\mathbf{r}_i),
\end{aligned}
$$

where $\mu(\cdot)$ denotes the dominating measure. Throughout, we let $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i)$ be completely unspecified. To estimate $\boldsymbol{\beta}$ without needing to specify this distribution, we take a semiparametric approach as described next.

1.2.2 Consistent and efficient estimator

Our approach is rooted in treating $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z})$ as an infinite dimensional nuisance parameter and using semiparametric techniques (Tsiatis 2006) to estimate $\boldsymbol{\beta}$. The approach involves first deriving the space spanned by this infinite-dimensional nuisance parameter. This space, referred to as the nuisance tangent space, and its orthogonal complement are derived in a similar way as Section S1 of Garcia & Ma (2016). The orthogonal complement of the nuisance tangent space serves as an intermediate calculation for the estimator of interest. Specifically, the efficient score function for $\boldsymbol{\beta}$, denoted $\mathbf{S}_{eff}$, is obtained by projecting the score function with respect to $\boldsymbol{\beta}$,

$$\begin{aligned} \mathbf{S}_{\beta}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) &\equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log\{f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta})\} \\ &= E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log\{f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta})\} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\right], \end{aligned} \tag{2}$$

onto the orthogonal complement of the nuisance tangent space. That is,

$$\mathbf{S}_{eff}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}) = \mathbf{S}_{\beta}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) - E\{\mathbf{h}(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\},$$

where $\mathbf{h}$ is a p-dimensional function that satisfies

$$E\{\mathbf{S}_{\beta}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{R}, \mathbf{X}, \mathbf{Z}\} = E[E\{\mathbf{h}(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\} \mid \mathbf{R}, \mathbf{X}, \mathbf{Z}].$$

To allow exchanging integration and differentiation in Equation (2), we assume that $f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta})$ and its partial derivative $\partial f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{Y} \mid \mathbf{R}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ are continuous functions of $\boldsymbol{\beta}$ and $\mathbf{R}$. The practical implementation of the procedure described above is however infeasible, because we are unable to perform the above

computation without the true distribution form of the random effect. To this end, we adopt a working model for $f_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$, denoted $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$, and perform the above calculation under such a working model. We provide the detailed expressions below, with all the affected quantities marked with *. Under such a working model, the score function with respect to $\boldsymbol{\beta}$ is

$$\mathbf{S}^*_{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) = E^*\left[\frac{\partial}{\partial\boldsymbol{\beta}}\log\{f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{Y}\mid\mathbf{R},\mathbf{X},\mathbf{Z},\boldsymbol{\beta})\}\mid\mathbf{Y},\mathbf{X},\mathbf{Z}\right], \tag{3}$$

and the locally efficient score function is

$$\mathbf{S}^*_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}) = \mathbf{S}^*_{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) - E^*\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z})\mid\mathbf{Y},\mathbf{X},\mathbf{Z}\},$$

Here, the "locally efficient score" means a function containing a working model in it. When a misspecified working model is used, the function has mean zero, and when a correct working model is used, the function is identical to the efficient score function, i.e. $\mathbf{S}_{eff}(Y,\mathbf{X},\mathbf{Z},\boldsymbol{\beta})$. An estimator based on solving the estimating equation formed by the locally efficient score function is named a locally efficient estimator. A locally efficient estimator subsequently has the property that if a misspecified working model is used, the estimator is consistent. When a correct working model is used, the estimator is efficient. Further, $\mathbf{h}^*$ is a p-dimensional function that satisfies

$$E\{\mathbf{S}^*_{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X},\mathbf{Z})\mid\mathbf{R},\mathbf{X},\mathbf{Z}\} = E[E^*\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z})\mid\mathbf{Y},\mathbf{X},\mathbf{Z}\}\mid\mathbf{R},\mathbf{X},\mathbf{Z}]. \tag{4}$$

An estimator of $\boldsymbol{\beta}$ is then obtained from solving the estimating equation

$$\sum_{i=1}^{n}\mathbf{S}^*_{eff}(\mathbf{y}_i,\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\beta}) = \mathbf{0}. \tag{5}$$

Using a working model $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ to replace the true form of $f_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ enables us to proceed with the computation. Of course, there is a cost involved with such a replacement. Fortunately, the cost is only in terms of estimation efficiency. The replacement does not affect the consistency of the resulting estimator.

Theorem 1. *The estimator $\widehat{\boldsymbol{\beta}}$ solving Equation (5) satisfies*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N\{0, \mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$$

in distribution when $n \to \infty$. Here, $\boldsymbol{\beta}_0$ is the true value of the parameter $\boldsymbol{\beta}$, $\mathbf{A} = E\left\{\partial \mathbf{S}^*_{eff}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}_0) \ /\partial \boldsymbol{\beta}^{\mathrm{T}}\right\}$, $\mathbf{B} = var\{\mathbf{S}^*_{eff}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}_0)\} = E\{\mathbf{S}^*_{eff}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}_0)^{\otimes 2}\}$. Additionally, if the true $f_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ is used in constructing the estimator, the resulting estimator $\widehat{\boldsymbol{\beta}}$ achieves the optimal estimation efficiency bound.

The proof of Theorem 1 is in the Appendix A. Theorem 1 implies that we are free to choose the form of $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ without incurring penalties on consistency or distorted type I error rates as in Heagerty & Kurland (2001), Agresti et al. (2004), Litiére et al. (2007, 2008). If $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ happens to be the true model, then the estimator for $\boldsymbol{\beta}$ achieves the optimal efficiency bound. For computational simplicity, we therefore choose the posited model of $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ as a standard normal distribution. In the simulation results given in Tables 1.5, 1.6, 1.7, 1.8, we show that even when the true dsitribution $f_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ is not from a standard normal, our estimator for $\boldsymbol{\beta}$ is still consistent. Regarding the estimation of the covariance matrix of our estimator, we estimate the derivative $\partial \mathbf{S}^*_{eff}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}^{\mathrm{T}}$ through numerical difference, and approximate the expectations via sample average.

A computational challenge in forming the estimating equation is solving Equation (4) for $\mathbf{h}^*$ as it is an ill-posed integral equation. However, as demonstrated next, a simple transformation of the response variable $\mathbf{Y}_{ij}$ and covariate $\mathbf{Z}_{ij}$ allows us to avoid solving this ill-posed problem.

1.2.3 Simplification of estimating equations

To circumvent the ill-posed problem in Equation (4), we transform the response variable $\mathbf{Y}_i$ and covariate $\mathbf{Z}_{ij}$ such that the transformed variables satisfy properties similar to the classical sufficiency and completeness.

Let $\mathbf{W}_i = \sum_{j=1}^{m} Y_{ij}\mathbf{Z}_{ij}$, $\mathbf{U}_i = (Y_{i(q+1)}, \ldots, Y_{im})^{\mathrm{T}}$. Write $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \cdots, \mathbf{Z}_{im}) = (\mathbf{Z}_{iL}, \mathbf{Z}_{iR})$, where $\mathbf{Z}_{iL} \in \mathcal{R}^{q\times q}$ and $\mathbf{Z}_{iR} \in \mathcal{R}^{q\times(m-q)}$. That is, $\mathbf{Z}_{iL}$ is the left $q \times q$ submatrix of $\mathbf{Z}_i$ and $\mathbf{Z}_{iR}$ is the right $q \times (m-q)$ submatrix of $\mathbf{Z}_i$. Let

$$\mathbf{M}_i = \begin{pmatrix} \mathbf{Z}_{iL} & \mathbf{Z}_{iR} \\ \mathbf{0}_{(m-q)\times q} & \mathbf{I}_{(m-q)\times(m-q)} \end{pmatrix}, \qquad \mathbf{M}_i^{-1} = \begin{pmatrix} \mathbf{Z}_{iL}^{-1} & -\mathbf{Z}_{iL}^{-1}\mathbf{Z}_{iR} \\ \mathbf{0}_{(m-q)\times q} & \mathbf{I}_{(m-q)\times(m-q)} \end{pmatrix}.$$

Under this notation, we transform $\mathbf{Y}_i$ as

$$\mathbf{Y}_i = \mathbf{M}_i^{-1} \begin{pmatrix} \mathbf{W}_i \\ \mathbf{U}_i \end{pmatrix}.$$

The matrix $\mathbf{M}_i$ is invertible because we assumed that the first q columns of $\mathbf{Z}_i$ form an invertible matrix. The one-to-one mapping from $(\mathbf{Z}_i, \mathbf{Y}_i)$ to $(\mathbf{W}_i, \mathbf{U}_i)$ allows us to take advantage of certain sufficiency and completeness properties of $\mathbf{W}_i$ and $\mathbf{U}_i$ as described in Theorem 2.

Theorem 2. *The variables* $\mathbf{W}_i$ *and* $\mathbf{U}_i$ *satisfy the following two properties:*

(a) Sufficiency of $\mathbf{W}$*:*

$$f_{\mathbf{U}|\mathbf{W},\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{w}, \mathbf{r}, \mathbf{x}, \mathbf{z}) = f_{\mathbf{U}|\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{w}, \mathbf{x}, \mathbf{z}) = f_{\mathbf{U}|\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{x}, \mathbf{z}),$$

$$f_{\mathbf{R}|\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{u}, \mathbf{w}, \mathbf{x}, \mathbf{z}) = f_{\mathbf{R}|\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{w}, \mathbf{x}, \mathbf{z}).$$

(b) Completeness of $\mathbf{W}$*:*

For any function $\mathbf{a}(\mathbf{w}, \mathbf{x}, \mathbf{z})$*, if* $E\{\mathbf{a}(\mathbf{W}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{R}, \mathbf{X}, \mathbf{Z}\} = \mathbf{0}$*, then* $\mathbf{a}(\mathbf{W}, \mathbf{X}, \mathbf{Z}) = \mathbf{0}$*.*

The proof of Theorem 2 is in Appendix. The sufficiency and completeness properties in Theorem 2 allow us to form a statistic free of the random slope associated with $\mathbf{Z}$ and remove the component containing the random slope from the estimating equation. Indeed Theorem 2 (a) yields that $E\{\mathbf{h}^*(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\}$ in Equation (4) is actually equal to $E\{\mathbf{h}^*(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{W}, \mathbf{X}, \mathbf{Z}\}$. The advantage of this equality is that the conditional expectation of $h^*(\mathbf{R}, \mathbf{X}, \mathbf{Z})$ given $(\mathbf{W}, \mathbf{X}, \mathbf{Z})$ satisfies

$$E^*\{\mathbf{h}^*(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{W}, \mathbf{X}, \mathbf{Z}\} \;=\; E\{\mathbf{S}^*_{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{W}, \mathbf{X}, \mathbf{Z}\}$$

and $E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\}$ has a closed form, given by

$$\frac{\sum_{\mathbf{u}} \mathbf{S}^*_\beta\{\mathbf{M}^{-1}(\mathbf{W}^{\mathrm{T}},\mathbf{u}^{\mathrm{T}})^{\mathrm{T}},\mathbf{X},\mathbf{Z}\}\exp\{(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta},\cdot,\mathbf{X}_q^{\mathrm{T}}\boldsymbol{\beta})(-\mathbf{Z}_L^{-1}\mathbf{Z}_R\mathbf{u}) + (\mathbf{X}_{(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\cdot,\mathbf{X}_m^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u}\}}{\sum_{\mathbf{u}}\exp\{(\mathbf{X}_1^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_q^{\mathrm{T}}\boldsymbol{\beta})(-\mathbf{Z}_L^{-1}\mathbf{Z}_R\mathbf{u}) + (\mathbf{X}_{(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_m^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u}\}}. \quad (6)$$

where the summation $\sum_{\mathbf{u}}$ is over all possible $\mathbf{u} \in \mathcal{R}^{m-p}$ such that each entry in $\mathbf{u}$ is either 0 or 1, and $\mathbf{u}$ satisfies that $\mathbf{W}_i = Z_{iL}\mathbf{Y}_{i1} + \mathbf{Z}_{iR}\mathbf{u}$. Here $\mathbf{Y}_{i1}$ is the subvector of $\mathbf{Y}_i$ formed by the first q elements.

Therefore, the estimating equation (5) which originally involved solving an ill-posed problem is now of the form

$$\sum_{i=1}^{n}\sum_{i=1}^{n}[\mathbf{S}^*_\beta(\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i) - E\{\mathbf{S}^*_\beta(\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i) \mid \mathbf{W}_i,\mathbf{X}_i,\mathbf{Z}_i\}] = \mathbf{0}. \quad (7)$$

All terms in the estimating equation can be explicitly constructed without needing to solve an ill-posed problem. The construction of $\mathbf{S}^*_\beta(\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i)$ does require specifying a proposed model $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$, but by Theorem 1, the model does not need to be correctly specified to ensure consistency. Therefore, we have constructed a simple estimation method that does not impose stringent assumptions on the unknown random effect, nor does it involve heavy computation. All terms in the new estimating equation are easy to compute with the most difficult part being $E\{\mathbf{S}^*_\beta(\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i) \mid \mathbf{W}_i,\mathbf{X}_i,\mathbf{Z}_i\}$, which we will use the Gaussian quadrature to deal with.

In summary, our algorithm for computing $\widehat{\boldsymbol{\beta}}$ involves:

Step 1. Specify a working model for $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$. For convenience, we suggest to model $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ using a normal distribution.

Step 2. Compute the function $\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z})$, as in Equation (3) where the expectations are computed under $f^*_{\mathbf{R}|\mathbf{X},\mathbf{Z}}$ from Step 1.

Step 3. Compute $E\{\mathbf{S}^*_\beta(\mathbf{Y}_i,\mathbf{X}_i,\mathbf{Z}_i) \mid \mathbf{W}_i,\mathbf{X}_i,\mathbf{Z}_i\}$ using Equation (6).

Step 4. Solve the estimating equation (7) to obtain $\widehat{\boldsymbol{\beta}}$.

1.3 Simulation study

1.3.1 Design of Simulation

We compared the performance of our estimator to the traditional normal-based maximum likelihood estimator (MLE). We used the *glmer* in R package *lme4* (Bates et al. 2016) to compute the maximum likelihood estimator. The assumption of the MLE is that the random effect is normally distributed, and that covariates and the random effects are independent. In comparison, our estimator does not assume that the random effect follows a specific distributional form, nor do we require independence between the covariates and the random effect. In this simulation study, we assess the sensitivity of our estimator and the MLE when these assumptions do not hold.

We generated 1000 data sets from the logistic random slope model in (1) with each data set having a sample size $n = 500$. We considered $m_i = 3$ covariates. We set the true parameter as $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\mathrm{T}}$. To assess the distributional assumption of the random effect, we generated data according to four different distributions:

1. Standard Normal random effect. R_i is from a standard normal distribution.

2. Mixed Normal random effect: R_i is from a mixture of normal distribution with 80% of the data from Normal(3,1), and 20% of the data from Normal(6,1.5).

3. Gamma random effect. R_i is from a Gamma distribution with shape parameter 1 and scale parameter 1.25.

4. Student-t random effect. R_i is from a student-t distribution with degree of freedom 3.

Thus, the distributional shapes of the random effect include the standard bell-shaped form, bimodality, heavy tailness and skewness. The deviations from the standard bell-

shaped form will allow us to assess how well our estimator performs in comparison to the MLE which assumes the random effect is indeed standard normal.

Under each of these four distributional assumptions for the random effect, we generated three different sets of covariates, first assuming their independence from the random effect:

1. Z_{ij} is from the Bernoulli distribution with success probability 0.5, and X_{ij} is from Normal(0.5,1);

2. Z_{ij} is from the Poisson distribution with parameter 0.5, and X_{ij} is from Normal(0.5,1);

3. Z_{ij} is from the Geometric distribution with success probability 0.7, and X_{ij} is from Normal(0.5,1).

Therefore, in total we considered 12 different cases: four ways of generating R_i's in combination with three ways of generating the covariates.

We further considered 12 additional cases similar to the above except that we introduced dependency between the random effect and covariates. This is aimed to assess deviations from the second assumption of the MLE in which the random effects and covariates are assumed to be independent. In the dependency case, we generated X_{ij} from Normal$(0.5R_i, 1)$ to achieve the dependency between R_i and X_{ij}'s. The generation of Z_{ij}'s were the same as before.

In summary, these settings were designed to investigate the performance of both the semiparametric estimator and the MLE when the random effect distribution is mis-specified, in combination with different covariate combinations of X and Z. For all data generation settings, we centered the generated random slopes to have zero mean to accommodate the standard normal-based MLE. In the proposed method, for all dependent and independent cases, we assumed the random effect is Normal(0,1)

distribution and is independent of all the covariates. This is of course not a valid assumption in all the settings considered above.

1.3.2 Simulation Results

We compared the performance of the semiparametric estimator and MLE in terms of their bias, sample variance, estimated variance, and 95% coverage probabilities. The results of the independent cases are given in Tables 1.1 to 1.4 and those for the dependent cases are given in Tables 1.5 to 1.8.

Tables 1.1 to 1.4 show that when covariates and random effects are independent, the semiparametric estimator has comparable performance to that of the MLE in terms of bias and the 95% coverage probabilities meeting the nominal level. While we expected the semiparametric estimator to be consistent based on Theorem 1, we were initially surprised by the robustness of MLE to deviations from normality. However, Neuhaus et al. (1992) demonstrated that the MLE actually performs quite well for mixed effect models when the random effect is not normally distributed. In terms of estimation variability, the semiparametric estimator has somewhat larger variability compared to the normal-based MLE, although the difference in variabilities is small. This is also within our expectation since MLE adopts stronger modeling assumptions and must have smaller estimation variability.

The results in Tables 1.5 to 1.8 indicate a different phenomenon. In the case when the covariates and random effect are dependent, inconsistency of the MLE starts to manifest. Specifically, the biases of the estimates from the normal-based MLE are sufficiently large, and they cause the coverage of the 95% confidence intervals to be completely off from the nominal level. In contrast, the biases of estimates from our proposed estimator is still very small, and the coverage probability of 95% confidence intervals remain close to their nominal level. This clearly demonstrates that if we treat the random effect as independent from the covariates while in fact there is

dependency between the two, the normal-based MLE loses its robustness and gives severely biased estimates with very small variability. Subsequently, inference based on MLE will be misleading. On the contrary, the semiparametric estimator continues to provide consistent estimation and valid inference results.

Summarizing the observations, the semiparametric estimator is a much more reliable method unless it is clear that the random effect and the covariates are independent of each other. Because the random effect is not observable, it is often difficult to determine its relation with the covariates. Thus, we recommend implementing the semiparametric estimator in general.

We also record the execution time of running 50 simulations using our estimator under one setting noted in Table 1.9. The CPU for this simulation is Intel I7-8700k@4.4GHz and the size of RAM is 32GB. From Table 1.9, the execution time increases as the cluster size increases or the number of parameters increases. These values show that the computation is generally sufficiently fast and we can use a single-thread R to run the entire simulation without engaging super computers with thousands of threads.

We also considered small sample performance, such as sample size $n = 25$ or $n = 50$. The algorithm does not converge in such sample sizes. As we gradually increase the sample size, we start to have some sensible results when the sample size $n = 220$. We report simulation results for $n = 220$ in Tables S.1 to S.8 in the supplement. Overall, the general conclusion based on $n = 220$ is the same as based on $n = 500$. The computing code is placed in the Supplementary Materials .

1.4 Analysis of a Huntington disease study

Huntington disease (HD) is a rare neurodegenerative disease linked to deterioration of the central nervous system. Its symptoms include unwanted choreatic movements, behavioral and psychiatric disturbances and dementia (Roos 2010). The

Cooperative Huntington Observational Research Trial (COHORT) was a large observational, longitudinal study of HD conducted from 2005 to 2011 that evaluated different cognitive and motor impairments associated with HD. The study included n=3211 participants who were annually evaluated over a four year time span. We focused on those subjects who had at least 4 consecutive visits during this study. Our main objective in analyzing COHORT is to investigate if cognitive measures are important determining the possibility of occurrence of HD. This objective stems from recent results that a major sign of HD is cognitive decline, and such decline can be observed long before motor symptoms first appear (Roos 2010).

To assess the association between cognitive measures and occurrence of HD, we modeled the data using the mixed effect logistic model in (1). For each person $i = 1, \ldots, n$, and visit $j = 1, \ldots, m$ with $m = 6$, we set the response variable Y_{ij} as 1 if the person was diagnosed with HD, and 0 otherwise. Diagnosis of HD occurs when the participant's extrapyramidal signs are unequivocally associated with HD and the diagnosis is determined by a trained clinician. We set Z_i to be the gender for subject i. We set $\mathbf{X}_{ij}$'s to be a set of four different motor and cognitive measures. Specifically, we set X_{1ij} to be the total motor score (TMS), defined as the sum of total motor impairments as evaluated using the Unified Huntington Disease Rating Scale (Group 1996). We set X_{2ij} to be the score from the Symbolic Digit Modality Test (SDMT), a test that assesses the cognitive impairment by some simple substitution tasks, such as visual scanning, attention, and motor speed. We set X_{3ij} to be the stroop color score (SCOLOR), a test that assesses the cognitive impairment by recording how many X's printed in blue, red, or green ink that a subject correctly verbally stated its color in a certain amount of time. We set X_{4ij} to be the stroop word score (SWORD), a test that assesses the cognitive impairment by recording the number of color words (blue, red, green) printed in black ink that a subject correctly verbally reads in a certain amount of time. We set X_{5ij} to be the stroop interference score (SINTER), a test

that assesses the cognitive impairment by recording how many color words that were printed in colored ink (eg. BLUE printed in green ink or BLUE printed in blue ink) and correctly verbally read by a subject in a certain amount of time. Lastly, we set R_i to be the random slope associated with Z_i.

We applied our proposed estimator and the standard-normal MLE to assess the association between cognitive impairments and occurrence of HD. We suspect that the cognitive covariates and random effect are dependent based on clinical results from Downing et al. (2008). They found gender differences in cognitive function. Females tended to outperform males on tests of memorization and language skills. Males tended to outperform females on tasks involving mathematical reasoning and visuospatial ability. These results suggest that if we assess the impact of cognitive measures on HD occurrence, we may have that cognitive measures and the random effect are dependent through gender. This would imply that the MLE could yield misleading results because it assumes independence, whereas our estimator does not.

We performed our analysis in two steps. In the first step, we analyzed three subsets of the data: the 1404 subjects who had four clinical visits, the 775 subjects who had five visits, and the 132 subjects who had six visits. In each of the three sub-data sets, we implemented the semiparametric estimator to obtain estimators $\widehat{\boldsymbol{\beta}}_1$, $\widehat{\boldsymbol{\beta}}_2$, $\widehat{\boldsymbol{\beta}}_3$, where $\widehat{\boldsymbol{\beta}}_1$=($\widehat{\beta}_{1tms}$, $\widehat{\beta}_{1sdmt}$, $\widehat{\beta}_{1scolor}$, $\widehat{\beta}_{1sword}$, $\widehat{\beta}_{1sinter}$), $\widehat{\boldsymbol{\beta}}_2$=($\widehat{\beta}_{2tms}$, $\widehat{\beta}_{2sdmt}$, $\widehat{\beta}_{2scolor}$, $\widehat{\beta}_{2sword}$, $\widehat{\beta}_{2sinter}$), $\widehat{\boldsymbol{\beta}}_3$=($\widehat{\beta}_{3tms}$, $\widehat{\beta}_{3sdmt}$, $\widehat{\beta}_{3scolor}$, $\widehat{\beta}_{3sword}$, $\widehat{\beta}_{4sinter}$). We then perform the second step by taking a weighted average of the results, i.e. we set $\widehat{\boldsymbol{\beta}}$=($\widehat{\beta}_{tms}$, $\widehat{\beta}_{sdmt}$, $\widehat{\beta}_{scolor}$, $\widehat{\beta}_{sword}$, $\widehat{\beta}_{sinter}$).

The weighted average is denoted as $\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{3} \mathbf{w}_i \widehat{\boldsymbol{\beta}}_i$, where the weights are proportional to the inverse of the variances of $\widehat{\boldsymbol{\beta}}_i$. That is, $\mathbf{w}_i$ is a diagonal matrix, with its jth element $w_{ij} = v_{ij}^{-1}/(\sum_{i=1}^{3} v_{ij}^{-1})$, where $v_{ij} = \text{var}(\widehat{\beta}_{ij})$. The variance of the final estimator is $\text{var}(\widehat{\beta}_j) = (\sum_{i=1}^{3} v_{ij}^{-1})^{-1}$. For comparison, we also implemented the normal-based MLE in the similar fashion.

Table 1.10 shows the results from both estimators. The semiparametric estimator indicates that cognitive scores from SDMT, SCOLOR, SWORD are not statistically significant, as their 95% confidence intervals contain zero. On the other hand, it detects TMS and SINTER to be significant covariates, both positively associated with the probability of developing HD. However, MLE indicates that only TMS score is statistically significant, while all the other four covariates are not statistically significant. The difference from the two analysis indicates that there is dependence between the random slope and the covariates. Based on both the theoretical results and the simulation experience, we believe the results from MLE can be misleading.

This result implies that if we adopt MLE on the data set to determine which covariates are needed for diagnosis of HD, we might neglect a vital covariate stroop interference score. The importance of stroop interference score coincided with clinical findings in (JS et al. 2013), where they found that prodromal HD patients have declined response shifting, and inhibition depends on efficient response shifting, while inhibition is necessary for stroop interference test. Based on these observations and our analysis results, we recommend using TMS and SINTER jointly to determine the occurrence of HD.

1.5 Discussion

We proposed a locally efficient estimator using a semiparametric approach in a mixed effect logistic model with random slope. Locally efficient means even when we use a misspecified working model, the resulting estimator of $\boldsymbol{\beta}$ is still consistent. If the true model happens to be the proposed working model, then the estimator is efficient. The method does not assume independence between the random slope and the covariates, and does not estimate or model the distribution of the random slope. In fact, an important advantage of the estimator is its consistency regardless whether or not the distribution of the random effect is correctly modeled, and regardless if

there is dependency between the random slopes and the covariates. Our method is developed under the mixed effect model with binary response under the logit link function. It will be interesting and valuable to investigate if the general approach can be adapted to incorporate the probit link or log-log link for the binary response, and to more general models in handling count or continuous response.

Sometimes, there is evidence that a random effect is discrete, hence it is natural to consider the treatment of a discrete random effect. In fact, if a random effect is discrete with infinitely many categories, we would recommend to ignore its discreteness and use a continuous working model for its distribution for computational purpose. In fact, our derivation has not assumed the random effect is continuous so the results derived before indeed apply. If a random effect is discrete with finitely many categories, the problem actually drastically simplifies. Indeed, in this case, treating the random effect probability masses as additional parameters, the original problem is a pure parametric model and a simple MLE will yield the efficient estimator.

Acknowledgement

Data from the COHORT study, which received support from HP Therapeutics, Inc., were used in this study. We thank the Huntington Study Group COHORT investigators and respective coordinators who collected the data, as well as participants and their families who made this work possible.

Yanyuan Ma is supported by a grant from the National Science Foundation.

Tanya P. Garcia is supported by a grant from the National Institute Of Neurological Disorders and Stroke of the National Institutes of Health.

The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Table 1.1: Simulation results when random effect and covariates are independent. Bias, sample variance (var), averaged estimated variance ($\widehat{\text{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\mathrm{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\text{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\text{var}}$	CI	bias	var	$\widehat{\text{var}}$	CI
	$R_i \sim 0.8N(3,1) + 0.2N(6,1.5)$ centered				$X_{ij} \sim \text{N}(0.5,1)$ $Z_i \sim \text{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	2.59	3.40	3.28	94.7%	0.44	0.45	0.44	94.6%
$\widehat{\beta}_2$	3.03	3.82	3.67	96.1%	0.48	0.49	0.50	96.2%
$\widehat{\beta}_3$	−2.94	3.76	3.38	94.4%	−0.38	0.43	0.45	95.5%
	$R_i \sim 0.8N(3,1) + 0.2N(6,1.5)$ centered				$X_{ij} \sim \text{N}(0.5,1)$ $Z_i \sim \text{P}oisson(0.5)$			
$\widehat{\beta}_1$	3.29	3.82	3.35	94.1%	0.21	0.45	0.43	94.4%
$\widehat{\beta}_2$	3.19	4.03	3.76	95.4%	0.25	0.48	0.49	95.4%
$\widehat{\beta}_3$	−2.62	3.93	3.45	94.1%	−0.30	0.44	0.44	95.9%
	$R_i \sim 0.8N(3,1) + 0.2N(6,1.5)$ centered				$X_{ij} \sim \text{N}(0.5,1)$ $Z_i \sim \text{G}eometric(0.7)$			
$\widehat{\beta}_1$	1.16	2.53	2.49	95.3%	0.078	0.41	0.41	95.6%
$\widehat{\beta}_2$	2.39	2.93	2.88	95.3%	0.13	0.47	0.47	94.0%
$\widehat{\beta}_3$	−1.59	2.74	2.58	95.3%	0.11	0.44	0.42	95.0%

Table 1.2: Simulation results when random effect and covariates are independent. Bias, sample variance (var), averaged estimated variance ($\widehat{\text{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\mathrm{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\text{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\text{var}}$	CI	bias	var	$\widehat{\text{var}}$	CI
	$R_i \sim \text{G}amma(1,1.25)$ centered				$X_{ij} \sim \text{N}(0.5,1)$ $Z_i \sim \text{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	2.17	3.34	3.09	95.4%	−2.33	0.38	0.39	92.8%
$\widehat{\beta}_2$	2.69	4.00	3.51	94.5%	−2.62	0.46	0.45	92.1%
$\widehat{\beta}_3$	−1.77	3.30	3.13	94.9%	−3.02	0.42	0.40	92.1%
	$R_i \sim \text{G}amma(1,1.25)$ centered				$X_{ij} \sim \text{N}(0.5,1)$ $Z_i \sim \text{P}oisson(0.5)$			
$\widehat{\beta}_1$	1.88	3.56	3.29	95.0%	−2.02	0.38	0.40	94.2%
$\widehat{\beta}_2$	1.90	4.05	3.78	94.5%	−2.61	0.45	0.45	93.1%
$\widehat{\beta}_3$	−1.24	3.46	3.38	95.1%	−2.91	0.41	0.41	93.8%
	$R_i \sim \text{G}amma(1,1.25)$ centered				$X_{ij} \sim \text{N}(0.5,1)$ $Z_i \sim \text{G}eometric(0.7)$			
$\widehat{\beta}_1$	1.82	2.58	2.43	95.4%	−2.00	0.37	0.38	94.2%
$\widehat{\beta}_2$	3.42	3.08	2.83	94.6%	−1.97	0.44	0.44	93.5%
$\widehat{\beta}_3$	−2.60	2.52	2.49	95.3%	−2.33	0.38	0.39	94.4%

Table 1.3: Simulation results when random effect and covariates are independent. Bias, sample variance (var), averaged estimated variance ($\widehat{\text{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\mathrm{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\text{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\text{var}}$	CI	bias	var	$\widehat{\text{var}}$	CI
	$R_i \sim$ t(3) centered				$X_{ij} \sim \mathrm{N}(0.5, 1)$ $Z_i \sim \mathrm{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	2.47	4.12	3.80	95.3%	−1.70	0.55	0.53	94.4%
$\widehat{\beta}_2$	5.51	5.21	4.44	93.4%	−1.22	0.60	0.60	94.3%
$\widehat{\beta}_3$	−3.61	4.39	3.91	95.4%	−2.10	0.54	0.55	94.3%
	$R_i \sim$ t(3) centered				$X_{ij} \sim \mathrm{N}(0.5, 1)$ $Z_i \sim \mathrm{P}oisson(0.5)$			
$\widehat{\beta}_1$	3.05	3.78	3.58	95.1%	−1.15	0.51	0.53	95.2%
$\widehat{\beta}_2$	4.29	4.61	4.11	94.4%	−0.90	0.59	0.60	95.4%
$\widehat{\beta}_3$	−2.73	3.87	3.67	95.0%	−1.70	0.52	0.54	94.8%
	$R_i \sim$ t(3) centered				$X_{ij} \sim \mathrm{N}(0.5, 1)$ $Z_i \sim \mathrm{G}eometric(0.7)$			
$\widehat{\beta}_1$	2.37	3.38	3.02	94.3%	−0.81	0.49	0.50	95.5%
$\widehat{\beta}_2$	2.77	4.00	3.47	94.2%	−1.22	0.57	0.56	93.5%
$\widehat{\beta}_3$	−2.07	3.43	3.07	94.3%	−1.28	0.53	0.51	94.6 %

Table 1.4: Simulation results when random effect and covariates are independent. Bias, sample variance (var), averaged estimated variance ($\widehat{\text{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\mathrm{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\text{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\text{var}}$	CI	bias	var	$\widehat{\text{var}}$	CI
	$R_i \sim \mathrm{N}(0, 1)$				$X_{ij} \sim \mathrm{N}(0.5, 1)$ $Z_i \sim \mathrm{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	2.44	3.42	3.25	95.3%	0.57	0.49	0.49	95.5%
$\widehat{\beta}_2$	3.60	4.28	3.75	94.9%	0.70	0.51	0.55	96.0%
$\widehat{\beta}_3$	−3.41	3.20	3.39	96.0 %	−0.40	0.52	0.50	94.5%
	$R_i \sim \mathrm{N}(0, 1)$				$X_{ij} \sim \mathrm{N}(0.5, 1)$ $Z_i \sim \mathrm{P}oisson(0.5)$			
$\widehat{\beta}_1$	2.56	3.42	3.35	94.2%	0.25	0.47	0.47	95.4%
$\widehat{\beta}_2$	4.76	4.48	3.88	94.1%	0.60	0.51	0.53	95.8%
$\widehat{\beta}_3$	−3.46	3.32	3.47	95.7%	−0.64	0.50	0.48	94.6%
	$R_i \sim \mathrm{N}(0, 1)$				$X_{ij} \sim \mathrm{N}(0.5, 1)$ $Z_i \sim \mathrm{G}eometric(0.7)$			
$\widehat{\beta}_1$	1.99	3.28	2.94	95.2%	0.44	0.45	0.47	96.1%
$\widehat{\beta}_2$	3.96	3.76	3.36	94.8%	1.19	0.53	0.53	95.1%
$\widehat{\beta}_3$	−1.86	3.13	2.96	95.2%	−0.71	0.47	0.48	95.7 %

Table 1.5: Simulation results when random effect and covariates are dependent: $X_{ij} \sim \text{N}ormal(0.5R_i, 1)$. Bias, sample variance (var), averaged estimated variance ($\widehat{\text{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\text{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\text{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\text{var}}$	CI	bias	var	$\widehat{\text{var}}$	CI
	$R_i \sim 0.8N(3,1) + 0.2N(6,1.5)$ centered				$Z_i \sim \text{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	1.74	3.76	3.25	94.6%	25.91	0.48	0.44	2.3%
$\widehat{\beta}_2$	2.72	3.85	3.72	96.0%	25.12	0.49	0.51	5.1%
$\widehat{\beta}_3$	−3.98	3.81	3.39	94.4%	29.41	0.42	0.41	0.9%
	$R_i \sim 0.8N(3,1) + 0.2N(6,1.5)$ centered				$Z_i \sim \text{P}oisson(0.5)$			
$\widehat{\beta}_1$	2.49	3.39	3.31	96.2%	25.40	0.43	0.44	1.6%
$\widehat{\beta}_2$	3.44	3.89	3.73	95.2%	24.32	0.50	0.51	4.8%
$\widehat{\beta}_3$	−3.52	3.77	3.41	94.7%	28.68	0.41	0.41	0.8%
	$R_i \sim 0.8N(3,1) + 0.2N(6,1.5)$ centered				$Z_i \sim \text{G}eometric(0.7)$			
$\widehat{\beta}_1$	1.86	3.12	2.76	94.1%	23.68	0.44	0.42	3.9%
$\widehat{\beta}_2$	3.15	3.70	3.18	93.2%	23.02	0.52	0.49	7.4%
$\widehat{\beta}_3$	−2.32	2.89	2.84	95.5%	27.28	0.41	0.40	2.3%

Table 1.6: Simulation results when random effect and covariates are dependent: $X_{ij} \sim \text{N}ormal(0.5R_i, 1)$. Bias, sample variance (var), averaged estimated variance ($\widehat{\text{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\text{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\text{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\text{var}}$	CI	bias	var	$\widehat{\text{var}}$	CI
	$R_i \sim \text{G}amma(1, 1.25)$ centered				$Z_i \sim \text{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	2.39	2.86	2.82	95.4%	20.73	0.39	0.41	8.2%
$\widehat{\beta}_2$	2.95	3.65	3.22	94.7%	20.16	0.51	0.48	16.6%
$\widehat{\beta}_3$	−1.90	3.22	2.86	95.2%	23.35	0.40	0.39	5.0%
	$R_i \sim \text{G}amma(1, 1.25)$ centered				$Z_i \sim \text{P}oisson(0.5)$			
$\widehat{\beta}_1$	2.73	3.12	3.07	95.8%	20.31	0.38	0.41	9.1%
$\widehat{\beta}_2$	3.17	4.11	3.55	94.6%	19.67	0.49	0.48	17.5%
$\widehat{\beta}_3$	−2.18	3.44	3.18	94.4%	22.72	0.39	0.40	6.2%
	$R_i \sim \text{G}amma(1, 1.25)$ centered				$Z_i \sim \text{G}eometric(0.7)$			
$\widehat{\beta}_1$	2.50	2.65	2.63	95.9%	19.80	0.40	0.41	10.8%
$\widehat{\beta}_2$	3.05	2.94	3.00	95.5%	19.05	0.49	0.47	19.6%
$\widehat{\beta}_3$	−1.29	2.85	2.67	94.8%	21.40	0.39	0.39	8.2%

Table 1.7: Simulation results when random effect and covariates are dependent: $X_{ij} \sim \mathrm{N}ormal(0.5R_i, 1)$. Bias, sample variance (var), averaged estimated variance ($\widehat{\mathrm{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\mathrm{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\mathrm{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\mathrm{var}}$	CI	bias	var	$\widehat{\mathrm{var}}$	CI
	$R_i \sim$ t(3) centered				$Z_i \sim \mathrm{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	2.16	4.84	4.33	95.2%	35.62	0.44	0.48	0.0%
$\widehat{\beta}_2$	4.30	5.79	4.94	94.4%	33.46	0.54	0.57	0.2%
$\widehat{\beta}_3$	−4.51	4.98	4.47	93.7%	42.23	0.38	0.41	0.0%
	$R_i \sim$ t(3) centered				$Z_i \sim \mathrm{P}oisson(0.5)$			
$\widehat{\beta}_1$	3.00	4.73	4.21	94.8%	34.37	0.43	0.47	0.0%
$\widehat{\beta}_2$	4.61	5.92	4.93	95.1%	32.16	0.53	0.55	0.3%
$\widehat{\beta}_3$	−3.28	4.56	4.34	95.3%	41.67	0.40	0.40	0.0%
	$R_i \sim$ t(3) centered				$Z_i \sim \mathrm{G}eometric(0.7)$			
$\widehat{\beta}_1$	1.86	3.78	3.43	94.5%	33.05	0.41	0.45	0.0%
$\widehat{\beta}_2$	3.73	4.29	3.98	93.3%	30.87	0.49	0.53	0.4%
$\widehat{\beta}_3$	−2.89	4.02	3.52	93.9%	40.16	0.36	0.39	0.0%

Table 1.8: Simulation results when random effect and covariates are dependent: $X_{ij} \sim \mathrm{N}ormal(0.5R_i, 1)$. Bias, sample variance (var), averaged estimated variance ($\widehat{\mathrm{var}}$), and the empirical coverage percentage of the 95% confidence interval (CI) for the semiparametric estimator and the normal-based MLE are reported. The true parameter $\boldsymbol{\beta} = (0.35, 0.6, -0.4)^{\mathrm{T}}$. Results are based on 1000 simulations with $n = 500$, $m_i = 3$. Biases are multiplied by 100, var and $\widehat{\mathrm{var}}$ are multiplied by 1000.

	Semiparametric Estimator				Normal-based MLE			
	bias	var	$\widehat{\mathrm{var}}$	CI	bias	var	$\widehat{\mathrm{var}}$	CI
	$R_i \sim \mathrm{N}(0,1)$				$Z_i \sim \mathrm{B}ernoulli(0.5)$			
$\widehat{\beta}_1$	3.05	4.11	3.70	95.4%	30.25	0.47	0.46	0.2%
$\widehat{\beta}_2$	5.42	4.98	4.21	94.3%	28.77	0.56	0.53	1.3%
$\widehat{\beta}_3$	−2.52	3.74	3.71	95.6%	35.39	0.40	0.41	0.1%
	$R_i \sim \mathrm{N}(0,1)$				$Z_i \sim \mathrm{P}oisson(0.5)$			
$\widehat{\beta}_1$	2.04	4.20	3.66	92.7%	29.55	0.47	0.45	0.2%
$\widehat{\beta}_2$	4.81	4.72	4.22	94.2%	28.04	0.50	0.52	1.2%
$\widehat{\beta}_3$	−2.16	3.72	3.64	97.0%	34.73	0.39	0.41	0.1%
	$R_i \sim \mathrm{N}(0,1)$				$Z_i \sim \mathrm{G}eometric(0.7)$			
$\widehat{\beta}_1$	1.61	3.20	2.92	94.7%	28.30	0.46	0.44	0.8%
$\widehat{\beta}_2$	3.85	3.44	3.39	94.8%	26.55	0.51	0.51	2.8%
$\widehat{\beta}_3$	−1.12	3.20	2.99	93.8%	33.22	0.460	0.40	0.2%

Table 1.9: Execution time of 50 simulations using our estimator when random effect is generated from Normal(0,1), Z_{ij} is from the Geometric distribution with success probability 0.7, Independent case means $X_{ij} \sim \mathrm{N}(0.5, 1)$ and dependent case means $X_{ij} \sim \mathrm{N}(0.5R_i, 1)$. The unit of time is second, and m stands for the cluster size and p denotes the number of parameters to be estimated.

	Independent	Dependent
m=2,p=3	79.39	78.59
m=3,p=3	208.3	201.28
m=4,p=3	321.74	310.9
m=3,p=1	41.16	39.54
m=3,p=2	112.47	113.76
m=3,p=3	205.56	202.47

Table 1.10: Results from Huntington disease (HD) data analysis based on semi-parametric estimator and normal-based maximum likelihood estimator (MLE). est: Parameter estimate, SE: standard error, 95% CI: 95% Wald-Type confidence interval, $\widehat{\beta}_{tms}$: Coefficient for total motor score, $\widehat{\beta}_{sdmt}$: Coefficient for symbol Digit Modalities Test, $\widehat{\beta}_{scolor}$: Coefficient for stroop color score, $\widehat{\beta}_{sword}$: Coefficient for stroop word score , $\widehat{\beta}_{sinter}$: Coefficient for stroop interference score. SE are multiplied by 10.

	Semiparametric Estimator			Normal-based MLE		
	Est	SE	95% CI	Est	SE	95% CI
$\widehat{\beta}_{tms}$	0.133	0.012	(0.065, 0.201)	0.266	0.004	(0.229, 0.303)
$\widehat{\beta}_{sdmt}$	0.028	0.012	(−0.040, 0.097)	−0.029	0.004	(−0.066, 0.009)
$\widehat{\beta}_{scolor}$	0.008	0.014	(−0.066, 0.081)	−0.029	0.003	(−0.063, 0.006)
$\widehat{\beta}_{sword}$	0.009	0.004	(−0.032, 0.048)	−0.014	0.002	(−0.039, 0.012)
$\widehat{\beta}_{sinter}$	0.074	0.002	(0.043, 0.104)	−0.014	0.004	(−0.053, 0.024)

Chapter 2

Prediction Using Many Samples with Working Models Containing Partially Shared Parameters

2.1 Introduction

Prediction is often the goal in many statistical analysis. Based on a statistical model and existing data with both covariates and responses, i.e., the labeled data, the usual practice is to estimate the unknown components of the model and use the resulting completely known model to predict the response associated with a new set of covariates, i.e. the unlabeled data. This practice works well when the labeled data and the new, unlabeled data share the same relation between the response (label) and the covariates. When additional labeled data are available, whose dependence of the response and covariates does not necessarily obey the same statistical rule as the original data, we usually cannot make use of them because information carried in such data may not benefit our prediction purpose.

However, it is not uncommon that even when data generated from different scenarios follow different models, they could still share some common components. For example, two data sets may both follow linear regression who share a common covariate subset and its effect. Further, if the response of the second data set is masked out and instead, only a dichotomous variable indicating whether or not the response is positive is available, then we will have two models share common covariate effects,

where one model has continuous response while the other categorical. The familiar mixed effect model can be viewed as one particular example as well. In this case, each cluster can be viewed as samples from a population and observations in the same cluster follow the same distribution. However, different cluster share some common features, captured as the fixed effect which is the same across all clusters. It is then natural to borrow information from observations in other clusters even if the main purpose is doing prediction in one specific cluster.

Such consideration directly leads us to consider prediction using multiple models from heterogeneous populations. Specifically, consider independent data sets from N populations. Let the jth data set be $Y_{ji}, \mathbf{X}_{ji}, \mathbf{Z}_i^{[j]}$ for $i = 1, \ldots, n_j$, where $\mathbf{X}_{ji}, j = 1, \ldots, N, i = 1, \ldots, n_j$ are i.i.d. observations and $\mathbf{Z}_i^{[j]}, i = 1, \ldots, n_j$ are i.i.d observations, but for $j_1 \neq j_2$, $\mathbf{Z}_i^{[j_1]}$ and $\mathbf{Z}_i^{[j_2]}$ can be different covariates. We describe the dependence of the response Y_{ji} on the covariates $\mathbf{X}_{ji}, \mathbf{Z}_j^{[j]}$ with a conditional probability density function (pdf)

$$f_{Y_{ji}|\mathbf{X}_{ji},\mathbf{Z}_i^{[j]}}(y, \mathbf{x}, \mathbf{z}^{[j]}) = f_j(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + {\mathbf{z}^{[j]}}^{\mathrm{T}}\boldsymbol{\alpha}_j, y, \boldsymbol{\gamma}_j), \quad j = 1, \ldots, N, \quad i = 1, \ldots, n_j. \quad (1)$$

Thus, even though different functional forms f_j and different parameter values $\boldsymbol{\alpha}_j, \boldsymbol{\gamma}_j$ are allowed in different populations, the population specific models still share a common parameter $\boldsymbol{\beta}$. Our purpose is to predict the response for a given set of covariates $\mathbf{X}, \mathbf{Z}^{[1]}$. Here, without loss of generality, we consider prediction in the first population. To distinction the model associated with the first population and the ones associated with the remaining populations, we name the first model the main model, while the rest the helper models.

To perform prediction, a natural practice is to first estimate $\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1$ using samples from the first population, and then use the estimated parameter values and the functional form f_1 to form a prediction, for example, the mean calculated as $\int y f_1(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + {\mathbf{Z}^{[1]}}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_1, y, \widehat{\boldsymbol{\gamma}}_1) d\mu(y)$, where $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\gamma}}_1$ are the estimators of $\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1$ and $\mu(\cdot)$ is the probability measure of response. One may also think of improving the

quality of the parameter estimation hence the quality of the prediction by using maximum likelihood estimator (MLE) based on all the data. This will improving the $\boldsymbol{\beta}$ estimator, and hence may further improve the estimation of other parameters and the prediction itself. Such consideration is based on a simple fact that all the models are correct. As long as one model, whether it is the main model or one of the helper models, is misspecified, such practice runs the risk of giving too much trust to the possibly misspecified model and hence can cause deteriorated prediction performance.

We take a different approach to making use of the multiple samples through a fusion of model averaging and meta analysis. This approach has the advantage that it can achieve the best prediction in large samples even if the main model that the prediction is based on is misspecified. On the other hand, if the main model is correct, the procedure will automatically eliminate the influence from misspecified helper models and use only the correctly specified helper models to form prediction. We can think of the method as a kind of fusion learning method. Given that in practice it is not usually known whether any model is correctly or incorrectly specified, this is a desirable feature. In addition, our procedure also has the flexibility of incorporating population specific statistical model forms, population specific parameters, and even population specific response variable type. For example, we allow some populations to have continuous response while others to have categorical responses. Finally, our procedure does not require pooling different samples from different populations together in order to carry out the parameter estimation. This can be a very important advantage since in our era data size is easily too large to handle, and our procedure can thus be used to justify splitting the data first, performing estimation separately, and then pool the results together. It in fact prescribes a method to pool the results together in an optimal way in terms of prediction performance.

The fusion learning method proposed in the current paper is related to the frequentist model averaging framework. The main difference between our method and the

classic works on frequentist model averaging, such as Yang (2001), Hjort & Claeskens (2003), Hansen (2007), Liu (2015), Chen et al. (2018), Zhang et al. (2018), Mitra et al. (2019), and Zhang & Xia (2019), is that we adopt different models on different data sets, while existing model averaging methods consider different models belong to the same regression family on a single data set. As far as we know, the current paper is the first one where different models are adaptively fitted for different data sets, and then an averaging approach is applied to improve the prediction for a target quantity. Our work is also related to meta-analysis in that we also consider multiple data sets. The classic goal of meta-analysis is mainly in parameter estimation, here we expand the scope of meta analysis by aiming at prediction. Further, classic meta-analysis generally assumes all the models are correctly specified, here we no longer make such assumption.

2.2 Prediction Procedure

The prediction procedure we propose is very simple. Given the model described in (1) and the observations $\{Y_{ji}, \mathbf{X}_{ji}, \mathbf{Z}_i^{[j]}\}$ for $i = 1, \ldots, n_j$, $j = 1, \ldots, N$, to predict Y associated with $\mathbf{X}, \mathbf{Z}^{[1]}$ from the first population, we carry out the following procedure.

Algorithm

Step 1: For $j = 1, \ldots, N$, estimate $\boldsymbol{\beta}$, $\boldsymbol{\alpha}_j$ and $\boldsymbol{\gamma}_j$ using the observations $\{Y_{ji}, \mathbf{X}_{ji}, \mathbf{Z}_i^{[j]}\}$, where $i = 1, \ldots, n_j$ via, for example, MLE. Denote the estimators $\widehat{\boldsymbol{\beta}}_{[j]}$, $\widehat{\boldsymbol{\alpha}}_j$ and $\widehat{\boldsymbol{\gamma}}_j$.

Step 2: Form the jth prediction of Y using the estimator $\widehat{\boldsymbol{\beta}}_{[j]}$ and $\widehat{\boldsymbol{\alpha}}_1$, $\widehat{\boldsymbol{\gamma}}_1$ by assigning

$$\widehat{Y}_j = E(Y \mid \mathbf{X}, \mathbf{Z}^{[1]}; \widehat{\boldsymbol{\beta}}_{[j]}, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\gamma}}_1) = \int y f_1(\mathbf{X}^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{[j]} + {\mathbf{Z}^{[1]}}^{\mathrm{T}} \widehat{\boldsymbol{\alpha}}_1, y, \widehat{\boldsymbol{\gamma}}_1) d\mu(y).$$

Step 3: Combine $\widehat{Y}_j$'s to construct a function of $\mathbf{w}$ through

$$\widehat{Y}(\mathbf{w}) = \sum_{j=1}^{N} w_j \widehat{Y}_j, \tag{2}$$

where $\mathbf{w} = (w_1, \ldots, w_N)^{\mathrm{T}}$ are a vector of weights that satisfy $0 \leq w_j \leq 1$ for $j = 1, \ldots, N$ and $\sum_{j=1}^{N} w_j = 1$. Let the set of all such $\mathbf{w}$'s be $\mathcal{W}$.

Step 4: Construct a crossvalidation criterion to evaluate the prediction performance of any set of weight choices, where

$$CV(\mathbf{w}) \equiv \frac{1}{n_1} \sum_{i=1}^{n_1} \{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2,$$

where $\widehat{Y}_{1i}^{(-i)}(\mathbf{w})$ is calculated the same way as $\widehat{Y}(\mathbf{w})$ described above except that Y_{1i} is left out of the calculation, i.e. it is the leave-the-ith-observation-out prediction of Y_{1i} under weight choice $\mathbf{w}$.

Step 5: Select $\mathbf{w}$ by minimizing the crossvalidated average prediction error, i.e.

$$\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w} \in \mathcal{W}} CV(\mathbf{w}). \tag{3}$$

Step 6: Let the resulting prediction be $\widehat{Y}(\widehat{\mathbf{w}})$. We term it fusion learning prediction (FLP).

The algorithm described above is rooted from a very basic idea of finding $\widehat{Y}(\mathbf{w})$ so that the expected error squared, i.e. $\mathrm{E}[\{\widehat{Y}(\mathbf{w}) - Y\}^2]$, is minimized. Because we do not want to put our full trust on the model $f_1(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}^{[1]^{\mathrm{T}}}\boldsymbol{\alpha}_1, y, \boldsymbol{\gamma}_1)$, we approximate $\mathrm{E}[\{\widehat{Y}(\mathbf{w}) - Y\}^2]$ via a model free fashion through using sample average, while mimicking the procedure of obtaining $\widehat{Y}(\mathbf{w})$ through the leave-one-out procedure $\widehat{Y}_{1i}^{(-i)}(\mathbf{w})$. Indeed, it is not difficult to show that $CV(\mathbf{w})$ is an asymptotically unbiased estimator of $\mathrm{E}[\{\widehat{Y}(\mathbf{w}) - Y\}^2]$, the quantity we would like to minimize. We provide the detailed verification of this result in the Appendix B.1.

Using crossvalidation to determine weights has been used as stacking (H.Wolpert 1992, Breiman 1996). In these seminal papers, the predictions averaged were from the same data set, and they did not consider heterogeneous populations as we considered here. In addition, the predictions to be averaged in our problem are calculated based

on both the main model and the helper models, while they are typically based on only one model in stacking. Lastly, in the current paper, we rigorously prove the optimality and weight assignment properties for combining predictions, while in stacking, these properties have not been built as far as we know.

2.3 Theoretical properties

It is not surprising that the theoretical properties of the above procedure derived below depend on whether or not the main model is misspecified, since the main model is the critical factor we rely on to perform prediction. Interestingly, although our prediction procedure also incorporates all the helper models, whether or not none, one or more of the helper models are misspecified does not affect the validity of the theoretical results.

2.3.1 Theoretical properties under misspecified main model

To present the theoretical properties, we first formally define the risk function as $R(\mathbf{w}) \equiv \mathrm{E}[\{\widehat{Y}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2]$. Ideally, one would certainly aim at minimizing $R(\mathbf{w})$ with respect to $\mathbf{w}$ in the set $\mathcal{W}$. In Theorem 3, we show that this goal is essentially achieved by our procedure. Usually, the prediction risk function is $\mathrm{E}[\{(Y(\mathbf{w}) - Y\}^2]$. It is seen that $\mathrm{E}[\{(Y(\mathbf{w}) - Y\}^2] = \mathrm{E}[\{(Y(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2] + \mathrm{E}[\{(Y - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2]$, where the second component is not controllable. Thus, we aim at minimizing $R(\mathbf{w})$.

Before stating Theorem 3, we first describe some assumptions that are needed to formally establish the theoretical properties of our predictor. All limiting processes considered in this paper correspond to $\underline{n} \to \infty$ where $\underline{n} = \min_{1 \le j \le N} n_j$. We allow both $N \to \infty$ and N fixed.

Assumption 1. *[Regularity of estimators] There exist values* $\boldsymbol{\alpha}_j^\star$, $\boldsymbol{\gamma}_j^\star$ *and* $\boldsymbol{\beta}_{[j]}^\star$ *such that* $N^{-1/2}n_j^{1/2}(\widehat{\boldsymbol{\alpha}}_j - \boldsymbol{\alpha}_j^\star) = O_p(1)$, $N^{-1/2}n_j^{1/2}(\widehat{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j^\star) = O_p(1)$ *and* $N^{-1/2}n_j^{1/2}(\widehat{\boldsymbol{\beta}}_{[j]} - \boldsymbol{\beta}_{[j]}^\star) = O_p(1)$ *uniformly for* $j = 1, \ldots, N$.

Assumption 1 is a simple requirement on the regularity of the estimators being averaged. This excludes, for example, super efficient estimators. When a model is correct, the limiting parameter values $\boldsymbol{\alpha}_j^\star$, $\boldsymbol{\gamma}_j^\star$ and $\boldsymbol{\beta}_{[j]}^\star$ are naturally the true parameter values, while when a model is misspecified, these limiting parameter values also exist and usually satisfy certain properties depending on how the estimators are constructed. Indeed, when the estimator is MLE, these results have been rigorously established in White (1982) under his conditions A1-A6. Here, for simplicity, we write the result in the general case as an assumption. To accomodate uniform convergence under a possibly diverging N, we weaken the convergence rate assumption to $(n_j/N)^{-1/2}$.

We write $\widehat{\boldsymbol{\theta}} \equiv (\widehat{\boldsymbol{\beta}}_{[1]}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_{[N]}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}})$ and $\boldsymbol{\theta}^\star \equiv (\boldsymbol{\beta}^{\star\mathrm{T}}_{[1]}, \ldots, \boldsymbol{\beta}^{\star\mathrm{T}}_{[N]}, \boldsymbol{\alpha}_1^{\star\mathrm{T}}, \boldsymbol{\gamma}_1^{\star\mathrm{T}})$. We now define some limiting quantities when the estimators are replaced by their corresponding limits. Specifically, let $Y_j^\star \equiv \int y f_1(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_{[j]}^\star + \mathbf{Z}^{[1]^{\mathrm{T}}}\boldsymbol{\alpha}_1^\star, y, \boldsymbol{\gamma}_1^\star) d\mu(y)$ and let $Y^\star(\mathbf{w}) \equiv \sum_{j=1}^N w_j Y_j^\star$. Similarly, let $Y_{1i,j}^\star \equiv \int y f_1(\mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_{[j]}^\star + \mathbf{Z}_i^{[1]^{\mathrm{T}}}\boldsymbol{\alpha}_1^\star, y, \boldsymbol{\gamma}_1^\star) d\mu(y)$ and let $Y_{1i}^\star(\mathbf{w}) \equiv \sum_{j=1}^N w_j Y_{1i,j}^\star$. We write the "risk" calculated under the limiting parameter values, as $R^\star(\mathbf{w}) \equiv \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right]$, and let $\xi \equiv \inf_{\mathbf{w}\in\mathcal{W}} R^\star(\mathbf{w})$ be the minimum risk under the ideal weights if the limiting parameter values were known. To be explicit, we write $\mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})$ as $g_{\mathrm{true}}(\mathbf{X}, \mathbf{Z}^{[1]}; \boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1)$. Then, $\mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) = g_{\mathrm{true}}(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1)$. Under the main model, we write $Y_j^\star = g(\mathbf{X}, \mathbf{Z}^{[1]}; \boldsymbol{\beta}_{[j]}^\star, \boldsymbol{\alpha}_1^\star, \boldsymbol{\gamma}_1^\star)$, $Y_{1i,j}^\star = g(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \boldsymbol{\beta}_{[j]}^\star, \boldsymbol{\alpha}_1^\star, \boldsymbol{\gamma}_1^\star)$, and $\widehat{Y}_j = g(\mathbf{X}, \mathbf{Z}^{[1]}; \widehat{\boldsymbol{\beta}}_{[j]}, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\gamma}}_1)$. Let $\epsilon_i = Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})$.

Assumption 2. *The expectations* $E(\epsilon_i^4)$, $E\left\{g^4(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \boldsymbol{\beta}_{[j]}^\star, \boldsymbol{\alpha}_1^\star, \boldsymbol{\gamma}_1^\star)\right\}$ *and* $E\left\{g_{true}^4(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1)\right\}$ *exist.*

Assumption 3. *$g(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \overline{\boldsymbol{\beta}}_{[j]}, \overline{\boldsymbol{\alpha}}_1, \overline{\boldsymbol{\gamma}}_1)$ is differentiable with respect to $\overline{\boldsymbol{\theta}}_{[j]}$, where $\overline{\boldsymbol{\theta}}_{[j]} \equiv (\overline{\boldsymbol{\beta}}_{[j]}^{\mathrm{T}}, \overline{\boldsymbol{\alpha}}_1^{\mathrm{T}}, \overline{\boldsymbol{\gamma}}_1^{\mathrm{T}})^{\mathrm{T}}$ and for any $\overline{\boldsymbol{\theta}}_{[j]}$ in a local neighborhood of $(\boldsymbol{\beta}_{[j]}^{\star\mathrm{T}}, \boldsymbol{\alpha}_1^{\star\mathrm{T}}, \boldsymbol{\gamma}_1^{\star\mathrm{T}})^{\mathrm{T}}$, there exists a positive constant $\bar{c}_1$ such that for any $1 \leq j \leq N$,*

$$E\left|g(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \overline{\boldsymbol{\beta}}_{[j]}, \overline{\boldsymbol{\alpha}}_1, \overline{\boldsymbol{\gamma}}_1)\right| \leq \bar{c}_1 \tag{4}$$

and

$$E\left\|\frac{g(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \overline{\boldsymbol{\beta}}_{[j]}, \overline{\boldsymbol{\alpha}}_1, \overline{\boldsymbol{\gamma}}_1)}{\partial \overline{\boldsymbol{\theta}}_{[j]}}\right\| \leq \bar{c}_1.$$

Assumption 4. *The expectations $E\left\{\sup_{1\leq j\leq N}(\widehat{Y}_j - Y_j^\star)^2\right\}$ and $E\left[\sup_{1\leq j\leq N}\{Y_j^\star - E(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right]$ exist.*

Assumptions 2, 3 and 4 are technical conditions on the existence, differentiability, and boundedness of various moments, and are rather mild conditions. Taking linear model $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}^{[1]\mathrm{T}}\boldsymbol{\alpha} + \epsilon$ as an example, the sufficient conditions for Assumptions 2 and 3 are $\mathrm{E}(\epsilon^4)$, $\mathrm{E}(\mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_i^{[1]\mathrm{T}}\boldsymbol{\alpha}_1)^4$, $\mathrm{E}(\mathbf{X}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_{[j]}^\star + \mathbf{Z}_i^{[1]\mathrm{T}}\boldsymbol{\alpha}^\star)^4$, $\mathrm{E}(\mathbf{X}_{1i}^{\mathrm{T}}\overline{\boldsymbol{\beta}}_{[j]} + \mathbf{Z}_i^{[1]\mathrm{T}}\overline{\boldsymbol{\alpha}}_1)$ and $\mathrm{E}(\|\mathbf{X}_{1i}\|^2 + \|\mathbf{Z}_i^{[1]}\|^2)$ exist. Assumption 4 is also a moment boundedness condition. When N is finite, it simply requires the existence of $\mathrm{E}\left\{(\widehat{Y}_j - Y_j^\star)^2\right\}$ and $\mathrm{E}\left[\{Y_j^\star - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right]$ for all models. To handle a diverging N, we write the corresponding requirement in the expectation of the supreme form.

Assumption 5. *[Misspecification of the main model] $\xi >> (\underline{n}/N^3)^{-1/2}$.*

Assumption 5 is a critical condition. It essentially requires the main model to be sufficiently misspecified. To see this, we first consider the fixed N case. In such case, if the main model had been correct, then $\boldsymbol{\beta}_{[1]}^\star = \boldsymbol{\beta}$, $\boldsymbol{\alpha}_1^\star = \boldsymbol{\alpha}_1$ and $\boldsymbol{\gamma}_1^\star = \boldsymbol{\gamma}_1$.

Consequently,

$$\begin{aligned}\xi &= \inf_{\mathbf{w}\in\mathcal{W}} R^\star(\mathbf{w}) \\ &= \inf_{\mathbf{w}\in\mathcal{W}} \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right] \\ &\leq \mathrm{E}\left[\{Y_1^\star - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right] \\ &= 0,\end{aligned}$$

hence Assumption 5 is violated. On the other hand, if the main model is misspecified with $\xi \asymp 1$, then Assumption 5 is satisfied. Here $\underline{n}^{-1/2}$ is the rate of convergence of ξ to zero under the true model case, hence it serves as the threshold between the true and misspecified case. To accomodate a possibly diverging N, the threshold is adapted to $(\underline{n}/N^3)^{-1/2}$. Similar assumptions can be found in Equation (7) of Ando & Lin (2014) and Assumption 1 (e) of Liu et al. (2020).

Under the above assumptions, the procedure we described in Section 2.2 leads to the optimal weight choice, in that the risk of the prediction using the estimated weights is the same as the risk of the prediction using the best possible weights to the leading order. The result is stated in Theorem 3 with its proof given in the Appendix B.2.

Theorem 3. *If Assumptions 1-5 are satisfied, then*

$$\frac{R(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} R(\mathbf{w})} \to 1 \tag{5}$$

in probability.

2.3.2 Theoretical properties under correct main model

We now consider the case that the main model is correctly specified. Because the helper models can be correct or misspecified, we let $\mathcal{D}$ be the subset of $\{1, \ldots, N\}$ that consists of the indices of the correctly specified models. Obviously $1 \in \mathcal{D}$. Let $\mathcal{D}^c$ be the complement of $\mathcal{D}$. Write $\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_N)^{\mathrm{T}}$. Let $\tau \equiv \sum_{j\in\mathcal{D}} w_j$, $\widehat{\tau} \equiv \sum_{j\in\mathcal{D}} \widehat{w}_j$ and M be the cardinality of $\mathcal{D}^c$.

Assumption 6. $\inf_{\mathbf{w}\in\mathcal{W},\sum_{j\in\mathcal{D}} w_j=0} R^\star(\mathbf{w}) >> \underline{n}^{-1/2}N^{3/2}$.

In practice, when the model complexity does not increase with the sample size, the risk of averaging misspecified models is typically of constant order. We hence also write Assumption 7 as an alternative of Assumption 6.

Assumption 7. $\{\inf_{\mathbf{w}\in\mathcal{W},\sum_{j\in\mathcal{D}} w_j=0} R^\star(\mathbf{w})\}^{-1} = O(1)$.

Assumption 6 obviously has similarity with Assumption 5. It imposes the relation between $N, \underline{n}$ and the optimal risk if all weights are assigned to the misspecified model. For simplicity, consider the fixed N case. Assumption 6 implies that the risk based on the misspecified models only is much larger than $n^{-1/2}$. When the main model is misspecified, the optimal risk automatically will be sufficiently large hence we do not need to restrict the weight set in Assumption 5. However, when the main model is correctly specified, the deterioration of the risk has to result from the mis-estimation of the parameter $\boldsymbol{\beta}$, hence the correct models will need to be excluded for this purpose. This is why we have to restrict the weights assigned to the correct model to be zero in Assumption 6. Assumption 6 serves as a separation between the performance of the correctly specified model and the misspecified model. Using any correct model in combination with the correct main model, the resulting risk in the limit will be zero. On the other hand, using any misspecified model, even when combined with the correct main model, the resulting risk in the limit will be much larger than $\underline{n}^{-1/2}$, as required by the assumption. Thus, intuitively it is clear that none of the incorrect model will be chosen when we minimize the risk. When N diverges, Assumption 6 is typically satisfied if $N^{3/2} = o(\underline{n}^{1/2})$ and $\inf_{\mathbf{w}\in\mathcal{W},\sum_{j\in\mathcal{D}} w_j=0} R^\star(\mathbf{w}) \asymp 1$. We summarize the result in Theorem 4, with its proof in the Appendix B.3.

Theorem 4. *If Assumptions 1, 2, 3 and 6 are satisfied, then* $\widehat{\tau} \to 1$ *in probability.*

Theorem 4 is a kind of model selection consistency, in that the method will automatically exclude the misspecified models. We stress that such model selection

consistency is based on the correctness of the main model f_1, and it only has the ability to exclude the misspecified helper models. For the case that the main model itself is misspecified, selection consistency in Theorem 4 does not hold in general and only prediction optimality in Theorem 3 applies. Regardless a helper model is correct or not, the prediction always relies on the main model hence the correctness/misspecification of the main model dominates the performance of the prediction procedure.

In modern applications, big data arise very often and it is common to perform prediction in each section of the data separately and then combine the predictions by simple average; see, for example, Li et al. (2013) and Battey et al. (2018). The problem with this practice is that as long as one model is misspecified, which is often the case due to the heterogeneity and complexity of data, the prediction error of the simple average can deteriorate very quickly. However, using the fusion learning procedure described in Section 2.2, the prediction properties described in Theorems 3 or 4 are automatically guaranteed and the procedure will yield better performance than simple average. We summarize this property in Corollary 1 and provide a brief argument in the Appendix B.4.

Corollary 1. *Assume the number of models and the sample size satisfy* $\underline{n} >> N^4M^{-2} + N^7M^{-4}$*, where* M *is the number of misspecified models. Regardless the main model is correct or not, under Assumptions 1-5 and Assumption 7, the risk of the fusion learning procedure in Section 2.2 is smaller than the risk associated with simple average to the first order.*

In Appendix B.5, we further explore the variance of the averaged prediction $\widehat{Y}(\widehat{w})$ for both the misspecified and correct main model cases, and establish that the variance of FLP, i.e. $\text{var}\{\widehat{Y}(\widehat{w})\}$, converges to zero under suitable conditions. We note that for the future observation Y, the prediction variance $\text{var}\{\widehat{Y}(\widehat{w}) - Y\} = \text{var}\{\widehat{Y}(\widehat{w})\} +$

var(Y), thus, our prediction is optimal in the sense that the only variability is the inherent variability associated with the randomness of the future observation.

2.4 Simulation Examples

2.4.1 Simulation Designs

We first consider linear regression models. We generate data from

$$f_j(Y_{ji}, \mathbf{X}_{ji}, \mathbf{Z}_i^{[j]}, \boldsymbol{\beta}, \boldsymbol{\alpha}_j) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\{Y_{ji} - \mathbf{X}_{ji}^{\mathrm{T}}\boldsymbol{\beta} - (1, {\mathbf{Z}_i^{[j]}}^{\mathrm{T}})\boldsymbol{\alpha}_j\}^2/2\sigma^2}, \tag{6}$$

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, N$, where Y_{ji} is a continuous response variable, $\boldsymbol{\beta} = (0.5, 0.6, -0.61, -0.48)^{\mathrm{T}}$, $\sigma = 0.5$, and $(\mathbf{X}_{ji}^{\mathrm{T}}, {\mathbf{Z}_i^{[j]}}^{\mathrm{T}})^{\mathrm{T}}$ is generated from a 9-dimensional multivariate normal distribution with mean zero and correlation structure as AR(1) with correlation coefficient 0.5 and variance 4. We first set $N = 3$, $n_1 = 100$, $n_2 = 200$, $n_3 = 100$, $\boldsymbol{\alpha}_1 = (0.4, 0.6, 0.5, -0.30, -0.25)^{\mathrm{T}}$, $\boldsymbol{\alpha}_2 = (0.49, 0.08, 0.09, -0.04, -0.06, 2.5)^{\mathrm{T}}$, and $\boldsymbol{\alpha}_3 = (0.51, 0.07, 0.1, -0.05, -0.04)^{\mathrm{T}}$. Note that here $\boldsymbol{\alpha}_2$ is a 6 dimensional vector, while $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_3$ are 5 dimensional vectors. When we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_2$ of the helper model 2, we always omit the last component of ${\mathbf{Z}_i^{[2]}}^{\mathrm{T}}$, so the helper model 2 is misspecified. When fitting the main model 1 and the helper model 3, we do not omit any component, so the main model 1 and the helper model 3 is correctly specified. This is our Design C1.1 (C means the main model is correctly specified).

When generating data for Design C1.2, the only difference from C1.1 is that in C1.2, $n_1 = 500$, $n_2 = 400$ and $n_3 = 300$.

When generating data for Design M1.1 (M means that the main model is misspecified), the difference between C1.1 and M1.1 is that in addition to misspecifying the helper model 2 as in C1.1, we now also misspecify the main model 1. Specifically, now we let $\boldsymbol{\alpha}_1 = (0.4, 0.6, 0.5, -0.30, -0.25, 0.1)^{\mathrm{T}}$, but when we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_1$ in the main model 1, we omit the last component of ${\mathbf{Z}_i^{[1]}}^{\mathrm{T}}$. When the dimension of $\boldsymbol{\alpha}_j$ is changed, we also change the dimension of $(\mathbf{X}_{ji}^{\mathrm{T}}, {\mathbf{Z}_i^{[j]}}^{\mathrm{T}})^{\mathrm{T}}$ when generating the dataset.

For Design M1.2, the sample sizes are $n_1 = 500$, $n_2 = 400$ and $n_3 = 300$, and the other setting is the same as those of Design M1.1.

Next, we consider to increase the number of helper models. In Design C1.3, we still generate data from (6), but now we let $N = 7$, $n_1 = 100$, $n_2 = 200$, $n_3 = 100$, $n_4 = 200$, $n_5 = 100$, $n_6 = 200$, and $n_7 = 100$. While we keep $\boldsymbol{\beta}$ as $(0.5, 0.6, -0.61, -0.48)^{\mathrm{T}}$, we let $\boldsymbol{\alpha}_1 = (0.4, 0.6, 0.5, -0.30, -0.25)^{\mathrm{T}}$, $\boldsymbol{\alpha}_2 = (0.49, 0.08, 0.09, -0.04, -0.06, 2.5)^{\mathrm{T}}$, $\boldsymbol{\alpha}_3 = (0.51, 0.07, 0.1, -0.05, -0.04)^{\mathrm{T}}$, $\boldsymbol{\alpha}_4 = (0.02, 0.05, -0.03, -0.01, 2.5)^{\mathrm{T}}$, $\boldsymbol{\alpha}_5 = (0.03, -0.07, 0.06, 0.02)^{\mathrm{T}}$, $\boldsymbol{\alpha}_6 = (-0.85, 2.5)^{\mathrm{T}}$, $\boldsymbol{\alpha}_7 = -0.87$. Note that now $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_5$ are 5-dimensional vectors, but $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_4$ are 6-dimensional vectors. When we estimate $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}_2^{\mathrm{T}})^{\mathrm{T}}$ of the helper model 2 and $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}_4^{\mathrm{T}})^{\mathrm{T}}$ of the helper model 4, we omit the last component of $\mathbf{Z}_i^{[2]}$ and $\mathbf{Z}_i^{[4]}$ respectively, so the helper model 2 and the helper model 4 are misspecified. In addition, when we estimate $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\alpha}_6^{\mathrm{T}})^{\mathrm{T}}$ of the helper model 6, we omit the last component of $\mathbf{Z}_i^{[6]}$, thus the helper model 6 is misspecified as well. In summary, in Design C1.3, we have the main model correctly specified, three helper models correctly specified, and three helper models misspecified,

Regarding Design C1.4, the difference between C1.3 and C1.4 is that we increased the sample sizes to $n_1 = 500$, $n_2 = 400$, $n_3 = 300$, $n_4 = 400$, $n_5 = 300$, $n_6 = 400$, and $n_7 = 300$.

For Design M1.3, the difference between C1.3 and M1.3 is that we misspecify the main model 1 in the same way as we did in M1.1.

In Design M1.4, the difference between M1.3 and M1.4 is that we increase the sample sizes to $n_1 = 500$, $n_2 = 400$, $n_3 = 300$, $n_4 = 400$, $n_5 = 300$, $n_6 = 400$, and $n_7 = 300$.

We further consider logistic models. We generate data from

$$f_j(Y_{ji}, \mathbf{X}_{ji}, \mathbf{Z}_i^{[j]}, \boldsymbol{\beta}, \boldsymbol{\alpha}_j) = \frac{\exp[Y_{ji}\{\mathbf{X}_{ji}^{\mathrm{T}}\boldsymbol{\beta} + (1, \mathbf{Z}_i^{[j]^{\mathrm{T}}})^{\mathrm{T}}\boldsymbol{\alpha}_j\}]}{1 + \exp\{\mathbf{X}_{ji}^{\mathrm{T}}\boldsymbol{\beta} + (1, \mathbf{Z}_i^{[j]^{\mathrm{T}}})^{\mathrm{T}}\boldsymbol{\alpha}_j\}}, \tag{7}$$

for $i = 1, \cdots, n_j$ and $j = 1, \ldots, N$, where $Y_{ji} = 0$ or 1. Then Designs C2.1, C2.2, M2.1, M2.2, C2.3, C2.4, M2.3 and M2.4 are the same as Designs C1.1, C1.2, M1.1, M1.2, C1.3, C1.4, M1.3 and M1.4 respectively, except that instead of generating data from (6), now we generate data from (7).

In addition, to demonstrate the flexibility of our method, we consider mixed helper models under two different settings of the main model, which we name Designs C1.5, C1.6, M1.5, M1.6 and Designs C2.5, C2.6, M2.5, M2.6. Designs C1.5, C1.6, M1.5, M1.6 are the same as Designs C1.3, C1.4, M1.3, M1.4 respectively, except that for $j = 2, 3$, we generate data from (7) instead of (6). So the main model and four helper models are linear models and the other two helper models are logistic. Design C2.5, C2.6, M2.5, M2.6 are the same as Design C2.3, C2.4, M2.3, M2.4 respectively, except that for $j = 2, 3$, we generate data from (6) instead of (7). So the main model and four helper models are logistic and the other two helper models are linear.

2.4.2 Comparison Methods

For comparison, in addition to our FLP method, we also implement six additional methods. The first comparison method is the simple average method, where we follow the same procedure as in the proposed method, except that we use an equal weight $w_i = 1/N$, instead of using crossvalidation to select a set of weights. The second comparison method is named "MLE main" method, where we ignore all helper models, and simply perform the standard prediction incorporating the MLE estimated parameters from the main model. The third comparison method is named the "MLE all" method, where we estimate parameters by maximizing the composite log-likelihood

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) = \sum_{i=1}^{n} \sum_{j=1}^{N} \log\{f_j(Y_{ji}, \mathbf{X}_{ji}, \mathbf{Z}_i^{[j]}, \boldsymbol{\beta}, \boldsymbol{\alpha}_j)\}.$$

Note that in the "MLE all" method, we use the helper models in the corresponding designs. The fourth and fifth comparison methods are respectively AIC and BIC

based model averaging methods, termed as smoothed AIC and smoothed BIC respectively (S. T. Buckland & Augustin 1997, Claeskens & Hjort 2008), where the averaging weight is respectively set as $\widehat{w}_j = \exp(-\text{AIC}_j/2)/\{\sum_j \exp(-\text{AIC}_j/2)\}$ and $\widehat{w}_j = \exp(-\text{BIC}_j/2)/\{\sum_j \exp(-\text{BIC}_j/2)\}$. Finally, the sixth method is a meta-analysis based prediction, where we first obtain the meta-analysis based estimator of $\boldsymbol{\beta}$, and then combine with $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\gamma}}_1$ to calculate the prediction for a new observation in the main population.

2.4.3 Simulation Results

We compare all methods using a test data set of 500 observations generated from the main model. Specifically, we calculate the mean prediction error $R(\mathbf{w}) = n_{new}^{-1} \sum_{i=1}^{n_{new}} \{\widehat{Y}_i(\mathbf{w}) - \text{E}(Y|\mathbf{X}_i, \mathbf{Z}_i^{[1]})\}^2$ on the testing data, where $n_{new} = 500$. We repeat the procedure for 100 times. The resulting average $R(\mathbf{w})$ values are provided in Tables 2.1 and 2.4. The "Gain" column in each table is the percentage of average $R(\mathbf{w})$ decreases when comparing our FLP method with one of the other six methods that has the smallest average $R(\mathbf{w})$.

In the last columns of Tables 2.1 and 2.6, we report the average $\widehat{\tau}$'s for the designs where the main model is correctly specified. It is seen that the average $\widehat{\tau}$'s are all above 0.8 except C2.5 in Table 2.6 and when the sample sizes are larger, they are closer to 1. This performance verifies the consistency shown in Theorem 4. The $\widehat{\tau}$ values of the designs with misspecified main models are not reported in the last columns of the tables, because there is no guarantee that the fusion learning procedure will tend to put specified weights on any models when the main model is misspecified.

When the main model is misspecified, from the results in the tables, we see that our FLP method yields the smallest average $R(\mathbf{w})$ compared to all the competitor methods, reflecting the theoretical result in Theorem 3. When the main model is correctly specified, our FLP method still has advantage. The possible reason is that

our FLP method asymptotically puts zero weights on the misspecified helper models, but the simple average, "MLE all", the smoothed AIC/BIC and the meta-analysis methods depend on the misspecified helper models substantially and the "MLE main" does not use any information from correctly specified helper models.

For smaller simple size or larger N, we find that the gain from our FLP method is more significant. We also find that the gain under logistic regression is more significant than that in linear regression.

Last, we inspect Tables 2.5 and 2.6. We can see that our FLP method still has obvious advantages over other methods. Similar to the previous findings, when sample size is smaller, the gain of our FLP method is more significant. For C-type designs, when the sample sizes are larger, the average τ's are closer to 1.

2.5 Real Data Example

We analyze the data "default of credit card clients" of an important bank in Taiwan, publicly available at the UCI machine learning repository. We consider six populations of clients, with credit scores equal to 10k, 110k, 210k, 310k, 410k, and $\geq 510k$ respectively. The sizes of the samples from the six populations are respectively 493, 588, 730, 272, 78 and 206.

The response variable is Y, with $Y_{ji} = 1$ if the ith client of the jth population defaulted, and Y_{ji}=0 otherwise. The covariates $X_{ji1}, \cdots, X_{ji5}$ are the payment ratios of the previous five months, which have numerical values between 0 and 1. The payment ratio is calculated as the payment amount of this month divided by the bill statement balance posted last month. If the bill statement balance is less than or equal to 0, or the payment amount is greater than the bill statement balance, then the payment ratio is set as 1. Other covariates include gender Z_{ji1}, with Z_{ji1}=1 representing male and Z_{ji1}=0 female, education level Z_{ij2}, with Z_{ji2}=1 indicating

university education or above and Z_{ji2}=0 otherwise, marital status Z_{ji3}, with Z_{ji3}=1 if married and Z_{ji3}=0 otherwise, and age Z_{ji4}.

We analyze the data set with the following logistic model:

$$f_j(Y_{ji}, \mathbf{X}_{ji}, \mathbf{Z}_i^{[j]}, \boldsymbol{\beta}, \boldsymbol{\alpha}_j) = \frac{\exp[Y_{ji}\{\mathbf{X}_{ji}^{\mathrm{T}}\boldsymbol{\beta} + (1, {\mathbf{Z}_i^{[j]}}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{\alpha}_j\}]}{1 + \exp\{\mathbf{X}_{ji}^{\mathrm{T}}\boldsymbol{\beta} + (1, {\mathbf{Z}_i^{[j]}}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{\alpha}_j\}}, \ j = 1, \cdot, 6, \ i = 1, \cdots, n_j. \quad (8)$$

We consider rotating the role of the six models. That is, each model will have one opportunity to serve as the main model while the other five models will serve as the helper models, so we have six settings. In the jth setting, we use the model associated with jth population as the main model. Table 2.7 shows the weights by our FLP method under all settings. It is interesting that in most settings, the weights of the helper models are larger than those of the main models, which means the helper models indeed help the predictions.

To check the performance of our FLP method, we randomly divide the data from the population of the main model into two parts with equal sizes. We repeat the procedure 100 times, and calculate the prediction errors for the methods compared in the simulation section, with prediction error calculated as

$$\textit{Prediction Error}_j = \frac{1}{100}\frac{1}{\lfloor n_j/2 \rfloor}\sum_{r=1}^{100}\sum_{i=1}^{\lfloor n_j/2 \rfloor}\left(\widehat{Y}_{ji}^{\{r\}} - Y_{ji}^{\{r\}}\right)^2, \ j = 1, \ldots, 6, \quad (9)$$

where $^{\{r\}}$ denotes the rth replication and n_j is the size of the sample from the jth population. Since

$$\mathrm{E}\left(\widehat{Y}_{ji} - Y_{ji}\right)^2 = \mathrm{E}\{\widehat{Y}_{ji} - \mathrm{E}(Y_{ji}|\mathbf{X}_{ji}, \mathbf{Z}_i^{[j]})\}^2 + \mathrm{E}(Y_{ji}^2) - \mathrm{E}^2(Y_{ji}),$$

the above prediction error also measures the prediction risk $\mathrm{E}\{\widehat{Y}_{ji} - \mathrm{E}(Y_{ji}|\mathbf{X}_{ji}, \mathbf{Z}_i^{[j]})\}^2$. Figure 2.1 shows the prediction errors under the six settings. First, we can see that our FLP method performs the most robust in all six settings. In the third, and last settings, our FLP method outperforms the simple average method and in the remaining four settings, they perform similarly. The "MLE main" method performs worse than FLP in all settings, possibly because it does not use information from

helper models. The "MLE all" method is also generally worse than FLP, where in the first, fifth and sixth settings, it performs much worse than FLP, and in the remaining, it performs similarly as FLP. The smoothed AIC and BIC methods have very similar performance, and they perform worse than FLP in the third, fifth and sixth settings, while have similar performance as FLP in the remaining settings. Finally, the meta-analysis method is in general worse than FLP as well, except in the second and fourth settings, where it performs similarly as FLP.

2.6 Concluding Remarks

In the context that a model of main research interest shares partial parameters with several other models, we have developed a fusion learning procedure to improve prediction for a new observation from the main model. The procedure achieves the optimal prediction risk when the main model is misspecified; and if the main model is correctly specified, the sum of weights assigned to the main model and the correct helper models converges to one. Numerical examples show that the procedure has excellent finite sample properties compared with the simple averaging method, the MLE using the main model only and the MLE using all models.

Let $Y_{1i,j}^{(-i)}$ be the prediction of Y_{1i} with Y_{1i} deleted using the jth model, $\mathbf{Y}_{1i}^{(-i)} = (Y_{1i,1}^{(-i)}, \ldots, Y_{1i,N}^{(-i)})^{\mathrm{T}}$, $\mathbf{l}$ be an $N \times 1$ vector with ones, and $\mathbf{H} = \sum_{i=1}^{n_1} (\mathbf{Y}_{1i}^{(-i)} - Y_{1i}\mathbf{l})(\mathbf{Y}_{1i}^{(-i)} - Y_{1i}\mathbf{l})^{\mathrm{T}}/n_1$. Then, by $\sum_{j=1}^{N} w_j = 1$, we have

$$\begin{aligned} CV(\mathbf{w}) &= \frac{1}{n_1}\sum_{i=1}^{n_1}\{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\{\mathbf{w}^{\mathrm{T}}(\mathbf{Y}_{1i}^{(-i)} - Y_{1i}\mathbf{l})\}^2 \\ &= \mathbf{w}^{\mathrm{T}}\frac{1}{n_1}\sum_{i=1}^{n_1}(\mathbf{Y}_{1i}^{(-i)} - Y_{1i}\mathbf{l})(\mathbf{Y}_{1i}^{(-i)} - Y_{1i}\mathbf{l})^{\mathrm{T}}\mathbf{w} = \mathbf{w}^{\mathrm{T}}\mathbf{H}\mathbf{w}, \end{aligned}$$

so minimizing $CV(\mathbf{w})$ is a quadratic programming and can be solved very quickly. As most of the literature on optimal model averaging, we did not study the limiting distribution of the resulting average prediction because of the difficulties produced

by the random weights and possible model misspecification. Using bootstrap may be a feasible way to solve this problem and this certainly warrants future work.

Table 2.1: $N = 3$ and all models are linear. FLP is the proposed method. Sim is the Simple Average Method. MLEm is the MLE main estimator. MLEa is the MLE all estimator. sAIC is the smoothed AIC method. sBIC is the smoothed BIC method. MT is the meta-analysis method. S and L in the Si column indicates the smaller and larger sample size in that design. D column indicates the specific simulation design in use.

		Average $R(\mathbf{w})\times 10^2$							Gain	Mean $\widehat{\tau}$
Si	D	FLP	Sim	MLEm	MLEa	sAIC	sBIC	MT		
S	C1.1	2.25	6.12	2.55	5.22	2.68	2.68	6.12	13.30%	0.96
L	C1.2	0.41	2.64	0.45	0.73	0.60	0.60	2.64	7.69%	0.97
S	M1.1	5.39	9.23	5.80	8.84	5.88	5.88	9.23	7.68%	–
L	M1.2	3.52	5.71	3.54	3.84	3.68	3.68	5.71	0.81%	–

Table 2.2: $N = 7$ and all models are linear. FLP is the proposed method. Sim is the Simple Average Method. MLEm is the MLE main estimator. MLEa is the MLE all estimator. sAIC is the smoothed AIC method. sBIC is the smoothed BIC method. MT is the meta-analysis method. S and L in the Si column indicates the smaller and larger sample size in that design. D column indicates the specific simulation design in use.

		Average $R(\mathbf{w})\times 10^2$							Gain	Mean $\widehat{\tau}$
Si	D	FLP	Sim	MLEm	MLEa	sAIC	sBIC	MT		
S	C1.3	2.10	4.30	2.50	2.92	2.48	2.47	4.30	18.02%	0.82
L	C1.4	0.39	1.71	0.45	0.42	0.60	0.62	1.71	9.41%	0.89
S	M1.3	5.37	7.42	5.88	6.12	5.67	5.64	7.42	5.04%	–
L	M1.4	3.44	4.74	3.51	3.55	3.65	3.66	4.74	2.11%	–

Table 2.3: $N = 3$ and all models are logistic. FLP is the proposed method. Sim is the Simple Average Method. MLEm is the MLE main estimator. MLEa is the MLE all estimator. sAIC is the smoothed AIC method. sBIC is the smoothed BIC method. MT is the meta-analysis method. S and L in the Si column indicates the smaller and larger sample size in that design. D column indicates the specific simulation design in use.

		Average $R(\mathbf{w})\times 10^2$							Gain	Mean $\widehat{\tau}$
Si	D	FLP	Sim	MLEm	MLEa	sAIC	sBIC	MT		
S	C2.1	1.57	1.82	1.93	3.53	2.05	2.05	1.89	15.74%	0.87
L	C2.2	0.33	0.72	0.36	1.13	0.58	0.58	0.64	9.50%	0.94
S	M2.1	1.60	1.81	1.93	3.48	2.11	2.11	1.86	13.18%	–
L	M2.2	0.39	0.76	0.42	1.15	0.64	0.64	0.68	7.30%	–

Table 2.4: $N = 7$ and all models are logistic. FLP is the proposed method. Sim is the Simple Average Method. MLEm is the MLE main estimator. MLEa is the MLE all estimator. sAIC is the smoothed AIC method. sBIC is the smoothed BIC method. MT is the meta-analysis method. S and L in the Si column indicates the smaller and larger sample size in that design. D column indicates the specific simulation design in use.

		Average $R(\mathbf{w})\times 10^2$							Gain	Mean $\widehat{\tau}$
Si	D	FLP	Sim	MLEm	MLEa	sAIC	sBIC	MT		
S	C2.3	1.39	2.04	1.83	3.77	2.01	1.90	2.00	32.11%	0.89
L	C2.4	0.29	1.02	0.33	1.69	0.58	0.56	0.89	16.16%	0.92
S	M2.3	1.58	2.32	2.00	4.06	2.26	2.15	2.31	26.39%	–
L	M2.4	0.37	1.09	0.42	1.83	0.65	0.63	0.97	14.58%	–

Table 2.5: $N = 7$, main model is linear, four helper models are linear and two helper models are logistic. FLP is the proposed method. Sim is the Simple Average Method. MLEm is the MLE main estimator. MLEa is the MLE all estimator. sAIC is the smoothed AIC method. sBIC is the smoothed BIC method. MT is the meta-analysis method. S and L in the Si column indicates the smaller and larger sample size in that design. D column indicates the specific simulation design in use.

		Average $R(\mathbf{w})\times 10^2$							Gain	Mean $\widehat{\tau}$
Si	D	FLP	Sim	MLEm	MLEa	sAIC	sBIC	MT		
S	C1.5	2.19	8.06	2.61	3.44	83.90	83.39	8.06	19%	0.83
L	C1.6	0.43	5.40	0.47	0.64	15.85	15.85	5.40	8%	0.91
S	M1.5	5.32	11.13	5.83	6.02	87.14	86.63	11.13	9%	–
L	M1.6	3.46	8.41	3.50	3.56	18.85	18.85	8.41	1%	–

Table 2.6: $N = 7$, main model is logistic, four helper models are logistic and two helper models are linear. FLP is the proposed method. Sim is the Simple Average Method. MLEm is the MLE main estimator. MLEa is the MLE all estimator. sAIC is the smoothed AIC method. sBIC is the smoothed BIC method. MT is the meta-analysis method. S and L in the Si column indicates the smaller and larger sample size in that design. D column indicates the specific simulation design in use.

		Average $R(\mathbf{w})\times 10^2$							Gain	Mean $\widehat{\tau}$
Si	D	FLP	Sim	MLEm	MLEa	sAIC	sBIC	MT		
S	C2.5	1.50	1.91	1.93	1.96	2.08	2.07	1.95	27.50%	0.76
L	C2.6	0.25	0.62	0.30	0.34	0.53	0.52	0.55	23.10%	0.83
S	M2.5	1.60	1.98	2.03	2.02	2.17	2.16	2.03	23.57%	–
L	M2.6	0.32	0.68	0.37	0.43	0.6	0.59	0.61	18.32%	–

Table 2.7: Weights obtained by our FLP method in the data analysis. In Setting j, the main model is Model j.

	Weights of the main and helper models					
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Setting 1	0.200	0.013	0.185	0.002	0.506	0.093
Setting 2	0.052	0.048	0.405	0.288	0.199	0.008
Setting 3	0.231	0.686	0.002	0.075	0.003	0.004
Setting 4	0.005	0.167	0.050	0.441	0.042	0.294
Setting 5	0.067	0.051	0.395	0.040	0.027	0.420
Setting 6	0.397	0.091	0.420	0	0.012	0.080

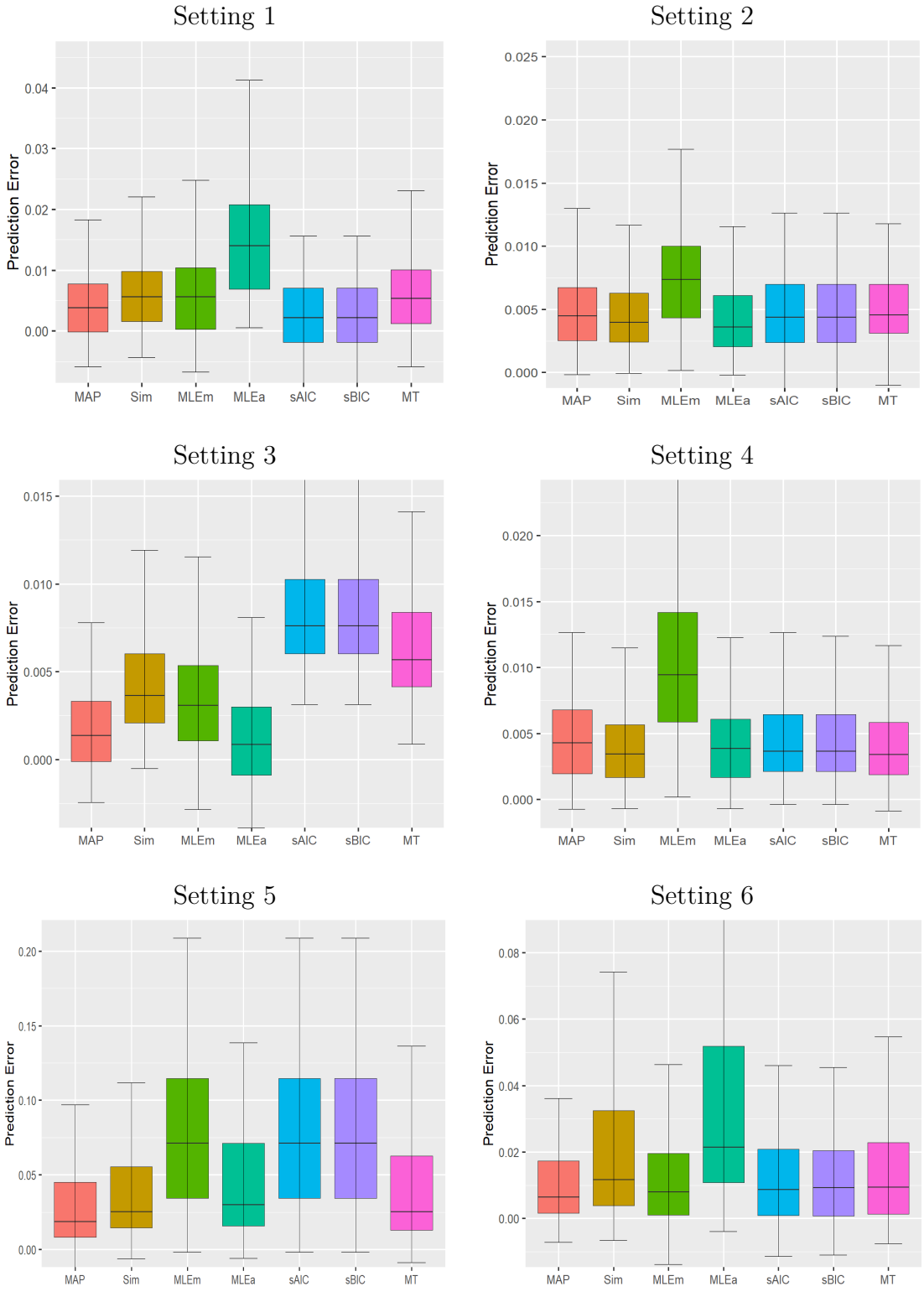

Figure 2.1: Boxplot of prediction errors in the real data example. In Setting j, the main model is Model j which is associated with the jth population.

Chapter 3

A Spline Assisted Pseudo-Likelihood Approach to Studying Binary Outcomes with Two-Phase Data.

3.1 Introduction

In general, for studying the association between a binary outcome and a set of covariates, it is common that data for some covariates can only be made available for a subset of subjects. The resultant incomplete data structure is usually described using a two phase sampling scheme (Neyman 1938, JE. 1982), where the outcome and the completely observed covariates are collected on all subjects at phase I and the remaining covariates are collected on a subset at phase II.

Inverse Probability Weighted estimator(IPW) requires a model for missingness mechanism and can be augmented to achieve double robustness. Pseudo-likelihood approach Breslow & Cain (1988) can be more efficient than IPW but requires a consistent estimate of missingness probability, therefore is most useful when phase I covariates are discrete so that a nonparametric estimate of the missingness probability is feasible. The existing MLEs either require that phase I covariates be discrete (Breslow & Holubkov 1997, Scott & Wild 1997), or need to limit the number of phase I covariates Tao et al. (2017).

3.2 Our Proposed Pseudo-Likelihood Method

Let Y denote the binary outcome variable with $Y = 1$ indicating cases and $Y = 0$ controls. Let $\mathbf{X}$ denote phase I covariates that are available for all subjects, and $\mathbf{Z}$ denote phase II covariates that are available only for a subset of subjects. It is of interest to fit a logistic regression model for describing the relationship between Y and all covariates $\mathbf{X}$ and $\mathbf{Z}$,

$$\text{logit pr}(Y = 1|\mathbf{X}, \mathbf{Z}) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_2, \tag{1}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the odd ratio(OR) parameters of interest. Note that, to simplify notation, a variable with value equal to one is implicitly included in $\mathbf{X}$, with the corresponding regression coefficient in $\boldsymbol{\beta}_1$ being the intercept parameter. Let $\boldsymbol{\beta}$ denote the vector of all parameters $(\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$.

3.2.1 Data and Setup

The observed data is described as follows. Let R be a binary variable indicating whether $\mathbf{Z}$ is observed or not. When $R = 1$, the complete data $(Y, \mathbf{X}, \mathbf{Z})$ is observed, but only $(Y, \mathbf{X})$ is observed when $R = 0$. The following scenarios are of interest:

(i) $(Y, \mathbf{X})$ is available for a cross-sectional sample of N subjects, and $\mathbf{Z}$ is available on a subset of m $(m \leq N)$ subjects; Missing completely at random: $p(R = 1|Y, \mathbf{X}, \mathbf{Z}) = \pi$ which is a constant;

(ii) $(Y, \mathbf{X})$ is available for a cross-sectional sample of N subjects, and $\mathbf{Z}$ is available on a subset of m $(m \leq N)$ subjects; Missing at random: $p(R = 1|Y, \mathbf{X}, \mathbf{Z}) = \pi(Y, \mathbf{X})$;

(iii) $(Y, \mathbf{X})$ is available for a case-control sample of N_1 cases and N_0 controls, $N = N_1 + N_0$. $\mathbf{Z}$ is available on a subset of m_1 $(m_1 \leq N_1)$ cases and m_0 $(m_0 \leq N_0)$ controls; $m = m_1 + m_0$. Missing completely at random: $p(R = 1|Y, \mathbf{X}, \mathbf{Z}) = \pi$

where π is a constant, or $p(R=1|Y=1,\mathbf{X},\mathbf{Z})=\pi_1$, $p(R=1|Y=0,\mathbf{X},\mathbf{Z})=\pi_0$, where π_1 and π_0 are constants and can be different;

(iv) $(Y,\mathbf{X})$ is available for a case-control sample of N_1 cases and N_0 controls, $N=N_1+N_0$. $\mathbf{Z}$ is available on a subset of m_1 ($m_1 \le N_1$) cases and m_0 ($m_0 \le N_0$) controls; $m=m_1+m_0$. Missing at random: $p(R=1|Y,\mathbf{X},\mathbf{Z})=\pi(Y,X)$.

Note that for cross-sectional sample, the problems in (i) and (ii) are the familiar missing covariate problem and are well studied. The efficient estimator is known to be the AIPW. In (iii), the $p(R=1|Y,\mathbf{X},\mathbf{Z})=\pi$ case is MCAR while the other two are actually MAR since the model depends on Y. In analyzing the problem, we can work with the general MAR model in (iv). Details on considerations of $\pi(Y,X)$ will be described below.

3.2.2 Our Idea of a Pseudo-Likelihood Method with Improved Efficiency

We proposed "goodness-of-fit" (GOF) and "balanced goodness-of-fit" (BGOF) two-phase sampling designs for selecting phase II subjects from a cross-sectional phase I sample, and found that BGOF improved efficiency for estimating odds ratio parameters compared to the case-control and balanced designs for both Phase I and Phase II covariates. These designs are built upon the difference between Y and its prediction based on an existing model linking Y and phase I variables $\mathbf{X}$, $P^e(Y=1|\mathbf{X})$, $S(y,\mathbf{x})=|y-P^e(Y=1|\mathbf{X})|$. For GOF, selected into phase II is based on a bernoulli experiment with success rate $\min\{1,c_1S(1,x_1)\}$ for a case and $\min\{1,c_0S(0,x_1)\}$ for a control, where c_1 and c_0 are two positive constants selected for achieving a targeted phase II sample size. BGOF performed sampling in two steps, generating a GOF sample first that is larger than the targeted sample size and adding an additional step of BD sampling to achieve the targeted sample size. To set the terminology, we

point out a simple relation

$$\begin{aligned} & \text{logit pr}(Y=1|\mathbf{X},\mathbf{Z},\mathbf{R}=1) \\ = \; & \log\{P(Y=1|\mathbf{X},\mathbf{Z},R=1)/P(Y=0|\mathbf{X},\mathbf{Z},R=1)\} \\ = \; & \text{logit pr}(Y=1|\mathbf{X},\mathbf{Z}) + \log\{P(R=1|Y=1,\mathbf{X})/P(R=1|Y=0,\mathbf{X})\}. \end{aligned}$$

Thus, the difference between logit $\text{pr}(Y=1|\mathbf{X},\mathbf{Z},\mathbf{R}=1)$ and logit $\text{pr}(Y=1|\mathbf{X},\mathbf{Z})$ is

$$o(\mathbf{x}) \equiv \log\{P(R=1|Y=1,\mathbf{X})/P(R=1|Y=0,\mathbf{X})\}, \tag{2}$$

which is termed an offset. This relation shows that the phase two data can be treated as if it is the whole data as long as we take into account the offset. For example, pseudo-likelihood methods were proposed to analyze GOF/BGOF data, where estimates are obtained by using standard logistic regression analysis software with inclusion of an offset term to adjust for sampling bias. For GOF, this offset term is written as

$$o(\mathbf{x}) = \log\ \min\{1, c_1 S(1,\mathbf{x})\} - \log\ \min\{1, c_0 S(0,\mathbf{x})\}.$$

For BGOF, the offset term equals $o(\mathbf{x}) + \log(\pi_{1l}/\pi_{0l})$, where π_{yl} is the fraction of subjects sampled from the GOF cell with $Y=y$ and stratum $L=l$.

This sampling design motivates the current work. Our rational for the superior efficiency of GOF/BGOF was that subjects lack-of-fit as measured by pearson residuals have a higher chance of being selected, and the lack-of-fit indicates that $\mathbf{Z}$ is needed in the external model to bring the subject back to fit. In this sense, data on $\mathbf{Z}$ for these subjects are more informative for association analysis. Without knowing how missingness for $\mathbf{Z}$ happened except for MCAR or MAR, we conjecture that by bringing into estimation information on the "lack-of-fit" of phase II subjects may help improve efficiency. Our proposed method is described as follows.

3.2.3 A Novel Pseudo-Likelihood Method when Phase I data is Case-Control and $\mathbf{Z}$ is missing MAR

(i) Fit a preliminary model between Y and $\mathbf{X}$ using phase I data with J strata, say logistic regression model $p^I(Y = 1|\mathbf{X}; \hat{\gamma})$. Superscript "I" indicates that phase I data was used to fit this model;

(ii) For cases, using the missingness indicator R as the outcome variable, fit a model

$$\text{logit pr}(R = 1|Y = 1, \mathbf{X}) = \boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_r + \sum_{j=1}^{J} \theta_{1j} I(\mathbf{X} \in S_j) + f_1\{p^I(Y = 0|\mathbf{X}; \hat{\gamma})\},$$

where $\mathbf{X}_r$ is a subset of $\mathbf{X}$ that is relevant to efficiency gain, which would often include rare exposures, S_j corresponds to the jth stratum, and f_1 is an unspecified smooth increasing function which will be estimated through B-spline approximation.

(iii) Similarly, for controls, using the missingness indicator R as the outcome variable, fit a model

$$\text{logit pr}(R = 1|Y = 0, \mathbf{X}) = \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{X}_r + \sum_{j=1}^{J} \theta_{0j} \mathbf{I}(\mathbf{X} \in S_j) + f_0\{p^I(Y = 1|\mathbf{X}; \hat{\gamma})\},$$

where $\mathbf{X}_r, S_j$ are the same as before, f_0 is an unspecified smooth increasing function which will be estimated through B-spline approximation.

Here, by including the strata information into our missingness model, we improved the efficiency.

(iv) The pseudo-likelihood estimating equation is written as

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} R_i(\mathbf{X}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}\{Y_i - \text{pr}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{R}_i = 1)\},$$

where

$$\begin{aligned}
&\text{logit pr}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{R}_i = 1) \\
= \;& \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + \log\{\text{expit}(\hat{\boldsymbol{\alpha}}_1^{\mathrm{T}}\mathbf{X}_{ri} + \sum_{j=1}^{J}\hat{\theta}_{1j}I(\mathbf{X}_i \in S_j) + \hat{f}_1\{p^I(Y_i = 0|\mathbf{X}_i;\hat{\gamma})\})\} \\
& -\log\{\text{expit}(\hat{\boldsymbol{\alpha}}_0^{\mathrm{T}}\mathbf{X}_{ri} + \sum_{j=1}^{J}\hat{\theta}_{0j}I(\mathbf{X}_i \in S_j) + \hat{f}_0\{p^I(Y_i = 1|\mathbf{X}_i;\hat{\gamma})\})\}
\end{aligned}$$

We will show that there exists a unique solution to the pseudo-likelihood estimating equation, the solution is consistent, and asymptotically normally distributed. We need to deal with (1) the internal estimation of γ parameters (in this sense, data is used twice) and (2) the estimated smoothers of f_1 and f_0.

3.3 Competing Approaches

We consider three competing methods implemented in the R package "OSdesign".

The first method, Maximum-Likelihood method(ML), maximizes the following pseudo-loglikehood function:

$$\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_{i=1}^{N} R_i \log \text{pr}(Y_i = y_i|\mathbf{X}_i, \mathbf{Z}_i, R_i = 1; \boldsymbol{\beta}) \\
&= \sum_{i=1}^{N} R_i \log \frac{\text{pr}(R_i = 1|Y_i = y_i, \mathbf{X}_i)\text{pr}(Y_i = y_i|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta})}{\sum_{y=0}^{1}\text{pr}(R_i = 1|Y_i = y, \mathbf{X}_i)\text{pr}(Y_i = y|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta})},
\end{aligned} \tag{3}$$

where

$$\text{pr}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{Z}_i)/\{1 + \exp(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{Z}_i)\}.$$

We approximate $\text{pr}(R_i = 1|Y_i = 0, \mathbf{X}_i)$ based on stratifying the data according to the covariates $\mathbf{X}_i$. Let the strata description be $\mathbf{X}_i \in S_j$, for $j = 1, \cdots, J$ and $i = 1, \cdots, N$. We approximate $\text{pr}(R_i = 1|Y_i = 0, \mathbf{X}_i, \mathbf{X}_i \in S_j)$ using μ_{0j}, where

$$\mu_{0j} = (n_{0j} - \gamma_{0j})/(N_{0j} - \gamma_{0j}),$$

and

$$\gamma_{0j} = n_{0j} - \sum_{i=1}^{N} R_i I(\mathbf{X}_i \in S_j)\text{pr}(Y_i = 0|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}).$$

Similarly, we also approximate $\text{pr}(R_i = 1|Y_i = 1, \mathbf{X}_i)$ based on stratifying the data according to the covariates $\mathbf{X}_i$. We approximate $\text{pr}(R_i = 1|Y_i = 1, \mathbf{X}_i, \mathbf{X}_i \in S_j)$ using μ_{1j}, where

$$\mu_{1j} = (n_{1j} - \gamma_{1j})/(N_{1j} - \gamma_{1j}),$$

and

$$\gamma_{1j} = n_{1j} - \sum_{i=1}^{N} R_i I(\mathbf{X}_i \in \mathbf{S}_j)\text{pr}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}).$$

Here, $n_{0j} = \sum_{i=1}^{N} I(Y_i = 0)I(\mathbf{X}_i \in S_j)R_i$, $n_{1j} = \sum_{i=1}^{N} I(Y_i = 1)I(\mathbf{X}_i \in S_j)R_i$. It is easy to see that γ_{0j} approximates $n_{0j} - (n_{0j} + n_{1j})\text{pr}(Y = 0 \mid R = 1, \mathbf{X}_i \in S_j)$, i.e. the difference of the number of observed $Y = 0$ (n_{0j}) and the expected $Y = 0$ ($(n_{0j} + n_{1j})\text{pr}(Y = 0 \mid R = 1, \mathbf{X}_i \in S_j)$) in stratum j of the complete sample. Similarly, γ_{1j} approximates $n_{1j} - (n_{0j} + n_{1j})\text{pr}(Y = 1 \mid R = 1, \mathbf{X}_i \in S_j)$, i.e. the difference of the number of observed $Y = 1$ (n_{1j}) and the expected $Y = 1$ ($(n_{0j} + n_{1j})\text{pr}(Y = 1 \mid R = 1, \mathbf{X}_i \in S_j)$) in stratum j of the complete sample. Thus, denoting $\mu_{kj} = (n_{kj} - \gamma_{kj})/(N_{kj} - \gamma_{kj})$ for $k = 0, 1$, $j = 1, \cdots, J$, the approximate function we maximize is written as

$$l(\boldsymbol{\beta}) \approx \sum_{i=1}^{N} R_i \text{log} \left\{ \sum_{j=1}^{J} \frac{I(\mathbf{X}_i \in S_j)\mu_{y_i j}\text{pr}(Y_i = y_i|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta})}{\mu_{1j}\text{pr}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) + \mu_{0j}\text{pr}(Y_i = 0|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta})} \right\}.$$

For implementation, we use iteration. Firstly, we set $\gamma_{0j}^{(0)} = \gamma_{1j}^{(0)} = 0$, hence $\mu_{0j}^{(0)} = n_{0j}/N_{0j}$, $\mu_{1j}^{(0)} = n_{1j}/N_{1j}$ and we maximize

$$L^{(0)}(\boldsymbol{\beta}) = \sum_{i=1}^{N} R_i \text{log} \left\{ \sum_{j=1}^{J} \frac{I(\mathbf{X}_i \in S_j)\mu_{y_i j}^{(0)}\text{pr}(Y_i = y_i|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta})}{\mu_{1j}^{(0)}\text{pr}(Y = 1|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta}) + \mu_{0j}^{(0)}\text{pr}(Y = 0|\mathbf{X}_i, \mathbf{Z}_i; \boldsymbol{\beta})} \right\}.$$

to obtain $\hat{\boldsymbol{\beta}}^{(0)}$. Then we plug $\hat{\boldsymbol{\beta}}^{(0)}$ to obtain

$$\gamma_{0j}^{(1)} = n_{0j} - \sum_{i=1}^{N} \text{pr}(Y_i = 0|\mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\beta}}^{(0)})R_i I(\mathbf{X}_i \in S_j),$$

$$\gamma_{1j}^{(1)} = n_{1j} - \sum_{i=1}^{N} \text{pr}(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i; \hat{\boldsymbol{\beta}}^{(0)})R_i I(\mathbf{X}_i \in S_j),$$

and form $\mu_{0j}^{(1)} = (n_{0j} - \gamma_{0j}^{(1)})/(N_{0j} - \gamma_{0j}^{(1)})$, $\mu_{1j}^{(1)} = (n_{1j} - \gamma_{1j}^{(1)})/(N_{1j} - \gamma_{1j}^{(1)})$. We iterate this process until the difference between $\gamma_{0j}^{(a+1)}$ and $\gamma_{0j}^{(a)}$, $\gamma_{1j}^{(a+1)}$ and $\gamma_{1j}^{(a)}$ is small enough to be considered as convergence for all $j = 1, \ldots, J$.

The second method is a Pseudo-Likelihood method(PL), where we aim at maximizing the same loglikelihood function in (3), except that we approximate the probabilitis $\text{pr}(R_i = 1 \mid Y_i, X_i)$ differently from ML. The approximation is based on stratify the data according to the covariates $\mathbf{X}_i$. Let the strata description be $\mathbf{X}_i \in S_j$, for $j = 1, \ldots, J$ and $i = 1, \ldots, N$. Based on

$$\begin{aligned}
\text{pr}(R_i = 1|Y_i = 1, \mathbf{X}_i \in S_j) &= \frac{\text{pr}(Y_i = 1, \mathbf{X}_i \in S_j|R_i = 1)\text{pr}(R_i = 1)}{\text{pr}(\mathbf{X}_i \in S_j \mid Y_i = 1)\text{pr}(Y_i = 1)} \\
&\approx \frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j)I(Y_i = 1)/n(n/N)}{\pi_1 \sum_{i=1}^{N} I(\mathbf{X}_i \in S_j)I(Y_i = 1)/\{\sum_{i=1}^{N} I(Y_i = 1)\}}
\end{aligned}$$

and

$$\begin{aligned}
\text{pr}(R_i = 1|Y_i = 0, \mathbf{X}_i \in S_j) &= \frac{\text{pr}(Y_i = 0, \mathbf{X}_i \in S_j|R_i = 1)\text{pr}(R_i = 1)}{\text{pr}(\mathbf{X}_i \in S_j, Y_i = 0)} \\
&\approx \frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j)I(Y_i = 0)/n(n/N)}{\pi_0 \sum_{i=1}^{N} I(\mathbf{X}_i \in S_j)I(Y_i = 0)/\{\sum_{i=1}^{N} I(Y_i = 0)\}},
\end{aligned}$$

we approximate the offset defined in (2) using

$$\begin{aligned}
& o(\mathbf{X}_i \in S_j) \\
\approx\ & \text{logpr}(R_i = 1|Y_i = 1, \mathbf{X}_i \in S_j) - \log \text{pr}(R_i = 1|Y_i = 0, \mathbf{X}_i \in S_j) \\
\approx\ & \log\left[\frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j)I(Y_i = 1)/n(n/N)}{\pi_1 \sum_{i=1}^{N} I(\mathbf{X}_i \in S_j)I(Y_i = 1)/\{\sum_{i=1}^{N} I(Y_i = 1)\}}\right] \\
& -\log\left[\frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j)I(Y_i = 0)/n(n/N)}{\pi_0 \sum_{i=1}^{N} I(\mathbf{X}_i \in S_j)I(Y_i = 0)/\{\sum_{i=1}^{N} I(Y_i = 0)\}}\right] \\
=\ & \log\{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j)I(Y_i = 1)\} - \log\{\sum_{i=1}^{N} I(\mathbf{X}_i \in S_j)I(Y_i = 1)\} \\
& +\log \sum_{i=1}^{N}\{I(Y_i = 1)\} - \log\pi_1 - \log\{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j)I(Y_i = 0)\} \\
& +\log\{\sum_{i=1}^{N} I(\mathbf{X}_i \in S_j)I(Y_i = 0)\} - \log \sum_{i=1}^{N}\{I(Y_i = 0)\} + \log\pi_0.
\end{aligned}$$

Note that $\pi_0 = 1 - \pi_1$, so we can view $o(X_i \in S_j)$ as a function of π_1, written as $o(X_i \in S_j, \pi_1)$. Thus, the likelihood in (3) is approximated as

$$l(\boldsymbol{\beta}) \approx \sum_{i=1}^{N} R_i \log \frac{\exp\{\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{Z}_i + o(\mathbf{X}_i \in S_j, \pi_1)\}}{\exp\{\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}_i + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{Z}_i + o(\mathbf{X}_i \in S_j, \pi_1)\} + 1}, \tag{4}$$

and we maximize it with respect to $\boldsymbol{\beta}, \pi_1$.

The third method, Weighted Likelihood method (WL), also known as Inverse Probability Weighting(IPW), uses the inverse probabilities $\mathrm{pr}(R_i = 1 \mid Y_i, \mathbf{X}_i)^{-1}$ to weigh the logistic regression loglikelihood functions computed based on the second stage data. Specifically, we maximize

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{R_i \log \mathrm{pr}(Y_i = 1 | \mathbf{X}_i, \mathbf{Z}_i)}{\mathrm{pr}(R_i = 1 | Y_i, \mathbf{X}_i)}, \tag{5}$$

where $\mathrm{pr}(Y_i = 1 | \mathbf{X}_i, \mathbf{Z}_i)$ is exactly the same as (1), to obtain WL estimate of $\boldsymbol{\beta}$. Here, to form the weights, we also stratify the data, and for $\mathbf{X}_i \in S_j$, we use

$$\mathrm{pr}(R_i = 1 \mid Y_i = 1, \mathbf{X}_i) \approx \frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j) I(Y_i = 1)}{\sum_{i=1}^{N} I(\mathbf{X}_i \in S_j) I(Y_i = 1)},$$
$$\mathrm{pr}(R_i = 1 \mid Y_i = 0, \mathbf{X}_i) \approx \frac{\sum_{i=1}^{n} I(\mathbf{X}_i \in S_j) I(Y_i = 0)}{\sum_{i=1}^{N} I(\mathbf{X}_i \in S_j) I(Y_i = 0)}.$$

3.4 Asymptotic Property of Our Estimator

In the following, we consider the population from which the phase I data is randomly extracted as the population of interest. Our goal is to estimate and make inference of $\widehat{\boldsymbol{\beta}}$ with respect to this population. Let $p^I(Y, \mathbf{X}, \boldsymbol{\gamma})$ be a working model of $\mathrm{pr}(Y \mid \mathbf{X})$ in the phase I data. Assumptions:

C1 There exists a $\boldsymbol{\gamma}^*$ so that $E\{\mathrm{pr}(Y \mid \mathbf{X})/p^I(Y, \mathbf{X}, \boldsymbol{\gamma}^*)\}$ is minimized.

C2 Among the diseased population, the true phase II sampling mechanism is $\mathrm{pr}(R = 1 | Y = 1, \mathbf{X}) = \mathrm{expit}[\mathbf{m}(\mathbf{X}, \boldsymbol{\tau}_1) + f_1\{p^I(0, \mathbf{X}, \boldsymbol{\gamma}^*)\}]$, and among the non-diseased population, the true phase II sampling mechanism is $\mathrm{pr}(R = 1 | Y = 0, \mathbf{X}) = \mathrm{expit}[\mathbf{m}(\mathbf{X})^{\mathrm{T}}\boldsymbol{\tau}_0 + f_0\{p^I(1, \mathbf{X}, \boldsymbol{\gamma}^*)\}]$. Here, $f_1(\cdot), f_0(\cdot)$ are smooth increasing functions with qth derivative.

C3 The spline order $r \geq q$.

C4 Denote the knots for $f_1(\cdot)$ as $t_{-r+1} = \cdots = t_0 = 0 < t_1 < \cdots < t_{M_1} < 1 = t_{M_1+1} = \cdots = t_{M_1+r}$, where M_1 is the number of interior knots and $[0,1]$ is divided into M_1+1 subintervals. M_1 satisfies $M_1 \to \infty$ and $M_1^{-1}N_1(\log N_1)^1 \to \infty$ and $M_1^{-1}N_1^{1/2} \to 0$ as $N_1 \to \infty$. Similarly, denote the knots for $f_0(\cdot)$ as $t_{-r+1} = \cdots = t_0 = 0 < t_1 < \cdots < t_{M_0} < 1 = t_{M_0+1} = \cdots = t_{M_0+r}$, where M_0 is the number of interior knots and $[0,1]$ is divided into M_0+1 subintervals. M_0 satisfies $M_0 \to \infty$ and $M_0^{-1}N_0(\log N_0)^{-1} \to \infty$ and $M_0^{-1}N_0^{1/2} \to 0$ as $N_0 \to \infty$.

C5 Let h_{1p} be the distance between the pth and $(p+1)$th interior knots. There exist two finite positive constants c, C so that $cM_1^{-1} \leq h_{1p} \leq CM_1^{-1}$ for all $p = r, \ldots, M_1 + r$. Similarly, let h_{0p} be the distance between the pth and $(p+1)$th interior knots. There exist two finite positive constants c, C so that $cM_0^{-1} \leq h_p \leq CM_0^{-1}$ for all $p = r, \ldots, M_0 + r$.

Theorem 5. *Assume the missingness model in (ii) and (iii) are correct. Then, under conditions C1-C5, as sample sizes $N_1, N_0 \to \infty$, the estimator $\widehat{\boldsymbol{\beta}}$ obtained from (iv) has the property that $N^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, \boldsymbol{\Sigma})$ in distribution. Here, $\boldsymbol{\Sigma}$ is defined this way:*

$$\begin{aligned}\boldsymbol{\Sigma} \;=\; & \mathbf{D}_\beta^{-1} E\left[\mathbf{U}_{ti}(\boldsymbol{\beta}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_0, \boldsymbol{\gamma}^*) + \mathbf{D}_\gamma \boldsymbol{\phi}_i(\boldsymbol{\gamma}^*)\right. \\ & -\mathbf{D}_{1,\zeta_1}\mathbf{A}_{\zeta_1}\left\{\mathbf{B}_{1,\zeta_1,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + I(Y_i=1)\mathbf{S}_{t1i}(\boldsymbol{\tau}_1, \boldsymbol{\gamma}^*)\right\} \\ & \left.-\mathbf{D}_{0,\zeta_0}\mathbf{A}_{\zeta_0}\left\{\mathbf{B}_{0,\zeta_0,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + I(Y_i=0)\mathbf{S}_{t0i}(\boldsymbol{\tau}_0, \boldsymbol{\gamma}^*)\right\}\right]^{\otimes 2} \mathbf{D}_\beta^{-1^{\mathrm{T}}}.\end{aligned}$$

It is seen that $\boldsymbol{\Sigma}$ has the typical sandwich form with several components. The components $\mathbf{D}_\beta$ and $\mathbf{U}_{ti}$ directly result from the estimating equation for $\boldsymbol{\beta}$. The component $\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*)$ is caused by including the working model $p^I(Y, \mathbf{X}, \boldsymbol{\gamma})$. If a completely determined working model had been used this term will vanish. The components $\mathbf{A}_{\zeta_1}$

and $\mathbf{S}_{t1i}$ are caused by the inclusion of the missingness model under $Y = 1$. Similarly for $\mathbf{A}_{\zeta_0}$ and $\mathbf{S}_{t0i}$. These terms will vanish if the covariate $\mathbf{Z}$ had been observed for all subjects, and the parameters $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_0$ will simplify to $\boldsymbol{\tau}_1, \boldsymbol{\tau}_0$ if we had not incorporated a spline approximation. The terms $\mathbf{B}_{1,\zeta_1,\gamma}$ and $\mathbf{B}_{0,\zeta_0,\gamma}$ capture the effect on the missingness estimation caused by estimating $\boldsymbol{\gamma}$. Finally $\mathbf{D}_{\gamma}$, $\mathbf{D}_{1,\zeta_1}$, $\mathbf{D}_{0,\zeta_0}$ capture the direct effect on the $\boldsymbol{\beta}$ estimation caused by estimating $\boldsymbol{\gamma}$ and the parameters in the missingness models. The detailed proof of Theorem 5 is in Appendix C.1.

Corollary 2. *If $f_1(\cdot)$ and $f_0(\cdot)$ are known instead of estimated via the spline approximation, the variance in Theorem 5 simplifies to*

$$\begin{aligned} \boldsymbol{\Sigma}_1 \;=\; & \mathbf{D}_{\beta}^{-1} E\left[\mathbf{U}_{ti}(\boldsymbol{\beta}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_0, \boldsymbol{\gamma}^*) + \mathbf{D}_{\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*)\right. \\ & -\mathbf{D}_{1,\tau_1}\mathbf{A}_{\boldsymbol{\tau}_1}\left\{\mathbf{B}_{1,\tau_1,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + I(Y_i = 1)[\mathbf{I}, \mathbf{0}]\mathbf{S}_{t1i}(\boldsymbol{\tau}_1, \boldsymbol{\gamma}^*)\right\} \\ & \left.-\mathbf{D}_{0,\tau_0}\mathbf{A}_{\boldsymbol{\tau}_0}\left\{\mathbf{B}_{0,\tau_0,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + I(Y_i = 0)[\mathbf{I}, \mathbf{0}]\mathbf{S}_{t0i}(\boldsymbol{\tau}_0, \boldsymbol{\gamma}^*)\right\}\right]^{\otimes 2} \mathbf{D}_{\beta}^{-1^{\mathrm{T}}}, \end{aligned} \tag{6}$$

where $\mathbf{D}_{1,\tau_1}$, $\mathbf{A}_{\boldsymbol{\tau}_1}$, $\mathbf{B}_{1,\tau_1,\gamma}$, $\mathbf{D}_{0,\tau_0}$, $\mathbf{A}_{\boldsymbol{\tau}_0}$, $\mathbf{B}_{0,\tau_0,\gamma}$ are the submatrices in $\mathbf{D}_{1,\zeta_1}$, $\mathbf{A}_{\boldsymbol{\zeta}_1}$, $\mathbf{B}_{0,\zeta_1,\gamma}$, $\mathbf{D}_{0,\zeta_0}$, $\mathbf{A}_{\boldsymbol{\zeta}_0}$, $\mathbf{B}_{0,\zeta_0,\gamma}$ that do not involve $\boldsymbol{\delta}_1$ or $\boldsymbol{\delta}_0$, and $[\mathbf{I}, \mathbf{0}]$ is a matrix conforming to the corresponding dimensions.

3.5 Simulation

To demonstrate the performance of our method, we conduct extensive simulations. In Simulation 1, our disease occurrence model is

$$\mathrm{pr}(Y_i = 1|\mathbf{X}_i, Z_i) = \exp\{(1, \mathbf{X}_i^{\mathrm{T}}, Z_i)\boldsymbol{\beta}\}/\{1 + \exp\{(1, \mathbf{X}_i^{\mathrm{T}}, Z_i)\boldsymbol{\beta}\}. \tag{7}$$

Here, $Y_i = 1$ means cases and $Y_i = 0$ means controls. $\mathbf{X}_i$ is a vector of length 7, whose first component X_{i1} has a Bernoulli distribution with success probability 0.1, and for the remaining components X_{i2} to X_{i7}, we first generate a intermediate quantity X_{temp} from a six-dimensional multivariate normal distribution with mean zero and correlation structure as AR(1) with correlation coefficient 0.5 and variance

1. Then we generate Z_i from a standard normal distribution. Lastly, we generate X_{i2} to X_{i7} using $0.05 \times Z_i + X_{temp}$. This data generation mechanism allows us to add dependency between X and Z, which is a common case in real data. We let the true $\boldsymbol{\beta}$ be $(-1.8, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55)^{\mathrm{T}}$. We use (7) to generate $N = 5000$ observations to form the phase I data, where we only retain $(\mathbf{X}_i, Y_i), i = 1, \ldots, N$ and masked out the Z_i's. We then select observations from the phase I data into the phase II data as the following. We considered the simple random sampling method (SRS) where the missingness model does not depend on Y_i and $\mathbf{X}_i$. Specifically, generated missingness indicator R_i's from a Bernoulli distribution with success probability 0.7. If $R_i = 1$, we retain the Z_i values hence the observation belongs to the phase II sample. Otherwise, if $R_i = 0$, we do not observe Z_i and the observation does not belong to the phase II sample. Thus, about 70% of the observations of the phase I sample is further selected into the phase II sample. The number of cases is about 24% in both phase I and phase II samples.

We implemented the proposed method as well as the competing methods ML, PL, WL. In the implementation of the ML, PL, WL methods, for both the phase I and phase II samples, we formed four strata, using the 0.25, 0.5 and 0.75 quantiles of X_2 from all the cases in the phase I samples. We report the bias and empirical standard deviation of the estimated $\widehat{\boldsymbol{\beta}}$ in Figure 3.1 and Table 3.1. When we inspect the empirical standard deviation of each individual parameter, we found that our method yields the smallest empirical standard deviation for 6 out of the 9 parameters. In contrast, ML wins for only two parameter, WL wins for only one parameter, and PL does not win in any parameter estimation. In addition, our method achieve the smallest overall mean squared error (0.507), which is about 8.6% smaller than the second smallest overall mean squared error (0.551) achieved by ML.

In Simulation 2, we use the Balanced Design (BD) to form phase II samples. We keep the total sample size $N = 5000$, and use the same setting to generate the

phase I sample. To determine a selection mechanism for generating the phase II sample, we split the phase I sample into four strata based on their X_2 value, where the separation points are the 0.25, 0.5 and 0.75 quartiles of X_2 from all the cases in the phase I samples. We then select all the cases in the phase I samples into the phase II data. We further count the number of cases in each stratum, denoted as $c_j, j = 1, \ldots, 4$, and we randomly select c_j controls from the jth stratum of the phase I control sample into the phase II sample, for $j = 1, \ldots, 4$. Likewise, we implemented the three competing methods in the same way as in Simulation 1 and the results are in Figure 3.2 and Table 3.2. After we exam the empirical standard deviation of each individual parameter, we found that our method yields the smallest empirical standard deviation for 4 out of the 9 parameters. Meanwhile, PL wins only for two parameter estimation, ML wins for three parameter estimation, and WL does not win any parameter estimation. Additionally, our method produced the smallest overall mean squared error (0.53), which is 15.3% smaller than the next smallest overall mean squared error (6.1) produced by PL.

In Simulation 3, we consider Case-Control Design (CC) to construct the phase II sample. The general simulation setting is the same as in Simulation 2, except that we do not form four strata based on $\mathbf{X}_2$. Specifically, once we generate the phase I sample, we denote the number of cases in the phase I samples as c_1. We then include all the c_1 cases of the phase I sample into the phase II sample, and we further randomly select c_1 controls from all the controls in the phase I sample with equal probability into the phase II sample. We analyze the data using the four competing methods as we did in simulation 1, and report the results in Figure 3.3 and Table 3.3. Reviewing the empirical standard deviation of each individual parameter, we find that our method produces the smallest empirical standard deviation for 3 out of the 9 parameters. As a comparison, PL wins for 3 parameters, ML wins for 3 parameters. WL does not win any parameters. We can also conclude that our method achieves

the smallest overall mean squared error (0.59), which is 14.8% smaller than the next smallest overall mean squared error (0.67) produced by PL.

In Simulation 4, we consider Goodness-of-Fit Based Design (GOF) for the phase II sample. The data generation is similar to Simulation 3, except that the selection probabilities of the controls into the phase II sample are no longer equal. To be specific, we generate the phase I sample in the same way as we did in Simulation 3. Let the number of cases in the phase I sample be c. We select all the c cases of the phase I sample into the phase II sample. To further select the controls into the phase II sample, we first adopt a working model

$$\text{logitpr}^e(Y_i = 1|\mathbf{X}_i, \boldsymbol{\eta}) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\eta} \tag{8}$$

to approximate the relation between Y_i and $\mathbf{X}_i$. To obtain $\boldsymbol{\eta}$, we generate a huge data set from model (7) and fit the data using model (8) to obtain $\widehat{\boldsymbol{\eta}}$. Among the controls of the phase I data, we define $S(\mathbf{x}_i) = \text{pr}^e(Y_i = 1|\mathbf{x}_i, \widehat{\boldsymbol{\eta}})$. We then select c observations from the control group with their selection probabilities proportional to $S(\mathbf{x}_i)$ to form the controls of the phase II sample. Results from the four methods are in Figure 3.4 and Table 3.4. In this case, our method yields the smallest empirical standard deviation for 8 out of 9 parameters. And only Pl wins for one parameters. WL and ML do not win any parameters. In the meantime, our method generated the smallest mean squared error (0.57), which is about 16.4% smaller than the next smallest mean squared error (0.66) generated by PL method.

In Simulation 5, we further consider Balanced Goodness-of-Fit Based Design (BGOF). We essentially adopt the Simulation 4 setting, except that the selection of the phase II controls are performed separately in two strata. To be specific, the phase I sample is generated the same as in Simulation 4, and all the cases in the phase I sample are selected into the phase II sample. We further compute $S(\mathbf{x}_i)$ as we did in Simulation 4. We then use the 0.5 sample quantile of all the X_{i2}'s in the phase I sample as the separation point to split the phase I sample into two strata. For

$j = 1, 2$, let c_j denote the number of cases in the jth stratum of the phase I sample. We select c_j observations from the controls of the jth stratum with their selection probabilities proportional to $S(\mathbf{x}_i)$ to form the controls of the phase II sample. The results of the analysis are in Figure 3.5 and Table 3.5. Here, our method leads to the smallest empirical standard deviation for 7 among the 9 parameters. The winner for the remaining parameter is ML. Not surprisingly, our method also achieves the smallest overall mean squared error (0.49), which is about 10.2% smaller than the next smallest mean squared error (0.54) provided by the PL method.

The results from Simulations 1 to 5 indicate that our proposed method is by far the most frequent winner in terms of estimation variability for different parameters, and it always achieves the smallest overall mean squared error among all methods. This is quite strong evidence of the superiority of our estimator. We also report the estimated standard deviation and 95% coverage probability of our method in the 'ESD' and 'CP' row respectively. We could see that the estimated standard deviation is very close to the empirical standard deviation, and the 95% coverage probability fluctuates around its nominal level. We thus recommend implementing our proposed method as a universally applicable method.

3.6 Concluding Remarks

In the context of modeling the binary outcome data set with incomplete data structure, we have developed a pseudo-likelihood approach to improve the efficiency for estimating the odds ratio parameters by incorporating the estimation information on the "lack-of-fit" of the phase II subjects. We have also derived its asymptotic properties. It has three major advantages over the existing methods dealing with the similar data set. The first advantage is that unlike the existing methods, it does not require to stratify the data set first if the phase I covariates are continuous, as the existing methods could only handle discrete phase I covariates. This is a desirable

feature as in the real world application, it is often difficult to determine how to stratify the data set properly for the continuous phase I covariates. The second advantage is that our method could also further improve the estimating efficiency via using the strata information if the data set could be stratified. The third advantage is that our method could handle relatively large number of continuous phase I covariates in a data set with relatively small sample size, as our method does not require stratification on that data set. In comparison, for the existing method, if the number of strata is too much, it is difficult to guarantee each strata will have enough samples for a valid estimation. The future work is to investigate how to handle high dimensional phase I covariates using the similar idea. When the phase I covariates $\mathbf{X}$ are high dimensional covariates, we can no longer use the logistic regression model as the preliminary model between the outcome variable $\mathbf{Y}$ and $\mathbf{X}$. Instead, we may consider use LASSO regression, hence it would be interested to study new approach under such high dimensional settings.

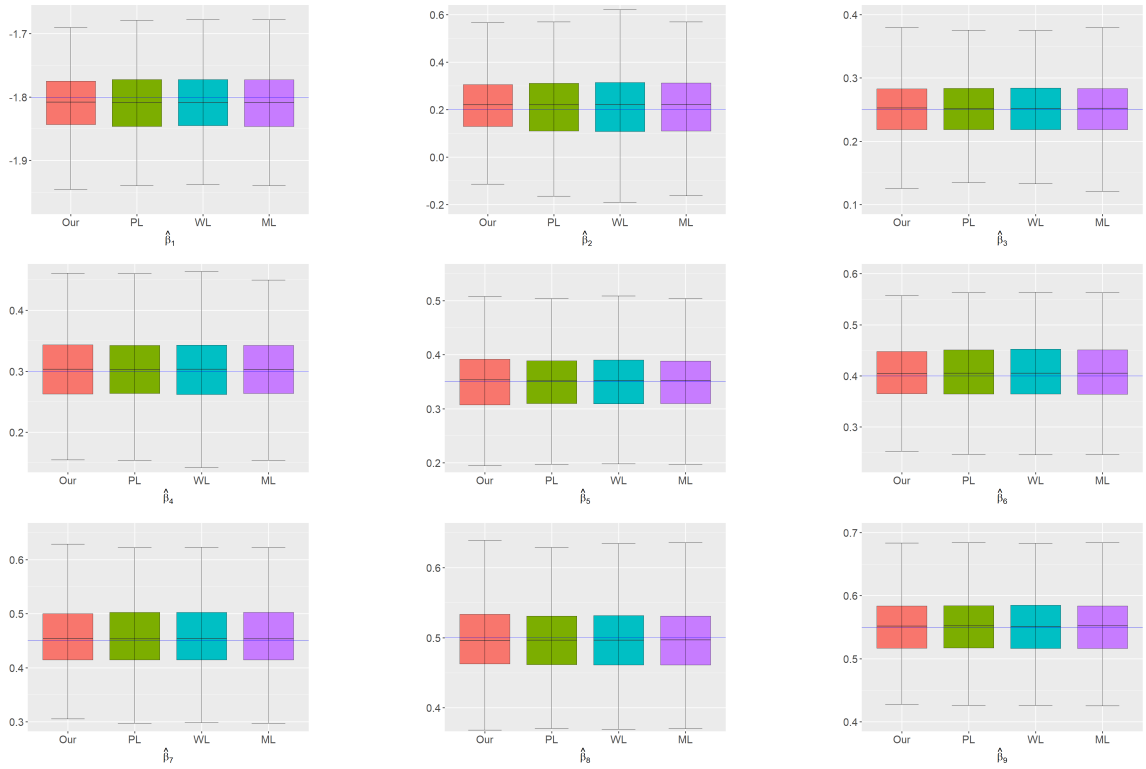

Figure 3.1: Boxplots of $\hat{\beta}_1$ to $\hat{\beta}_9$ using the four methods for the Simple Random Sampling phase II data in Simulation 1. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods . The blue horizontal line in each boxplot is the true parameter value.

Table 3.1: Empirical bias, empirical standard deviation of $\hat{\boldsymbol{\beta}}$ from four methods using the Simple Random Sampling phase II data in Simulation 1. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. All results are timed by 100 and rounded to 3 digits.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Ours	Bias	1.08	1.51	0.09	0.29	0.04	0.65	0.59	0.23	0.10
	SD	5.16	13.98	5.48	5.96	6.16	6.24	6.11	5.37	4.72
	ESD	5.35	13.50	5.44	6.09	6.11	6.12	6.14	5.53	4.88
	CP	96%	94%	94%	95%	95%	94%	95%	96%	96%
PL	Bias	1.06	1.48	0.06	0.33	0.01	0.68	0.64	0.17	0.10
	SD	5.60	15.16	5.15	5.99	6.15	6.34	6.21	5.49	4.72
WL	Bias	1.05	1.47	0.06	0.32	0.01	0.68	0.64	0.17	0.10
	SD	5.60	15.18	5.15	5.98	6.15	6.35	6.22	5.50	4.72
ML	Bias	1.05	1.48	0.06	0.33	0.01	0.68	0.64	0.17	0.10
	SD	5.59	15.16	5.13	5.99	6.15	6.34	6.21	5.49	4.71

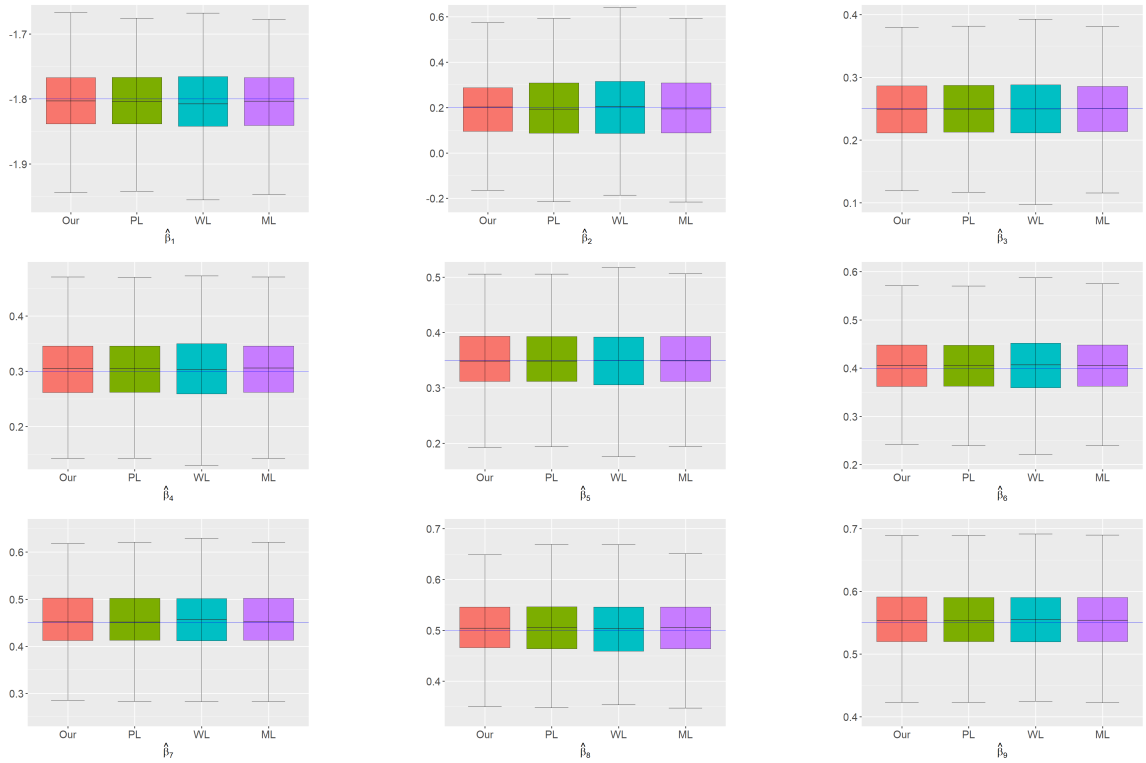

Figure 3.2: Boxplots of $\hat{\beta}_1$ to $\hat{\beta}_9$ using the four methods for the Balanced Design Sampling phase II data in Simulation 2. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. The blue horizontal line in each boxplot is the true parameter value.

Table 3.2: Empirical bias, empirical standard deviation of $\hat{\boldsymbol{\beta}}$ from four methods using the Balanced Design Sampling phase II data in Simulation 2. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. All results are timed by 100 and rounded to 3 digits.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Ours	Bias	0.33	0.48	0.11	0.36	0.03	0.63	0.38	0.56	0.60
	SD	5.58	13.82	5.32	6.27	6.25	6.33	6.63	6.08	5.09
	ESD	5.41	14.26	5.34	6.50	6.54	6.57	6.59	5.96	5.20
	CP	94%	95%	96%	96%	95%	95%	96%	95%	96%
PL	Bias	0.33	0.18	0.13	0.36	0.04	0.62	0.38	0.57	0.61
	SD	5.62	16.24	5.32	6.28	6.25	6.33	6.63	6.09	5.09
WL	Bias	0.55	0.01	0.18	0.36	0.09	0.78	0.58	0.50	0.59
	SD	5.84	16.94	5.62	6.77	6.51	6.57	6.96	6.42	5.34
ML	Bias	0.42	0.18	0.12	0.37	0.05	0.63	0.39	0.57	0.61
	SD	5.62	16.23	5.30	6.28	6.25	6.33	6.63	6.09	5.10

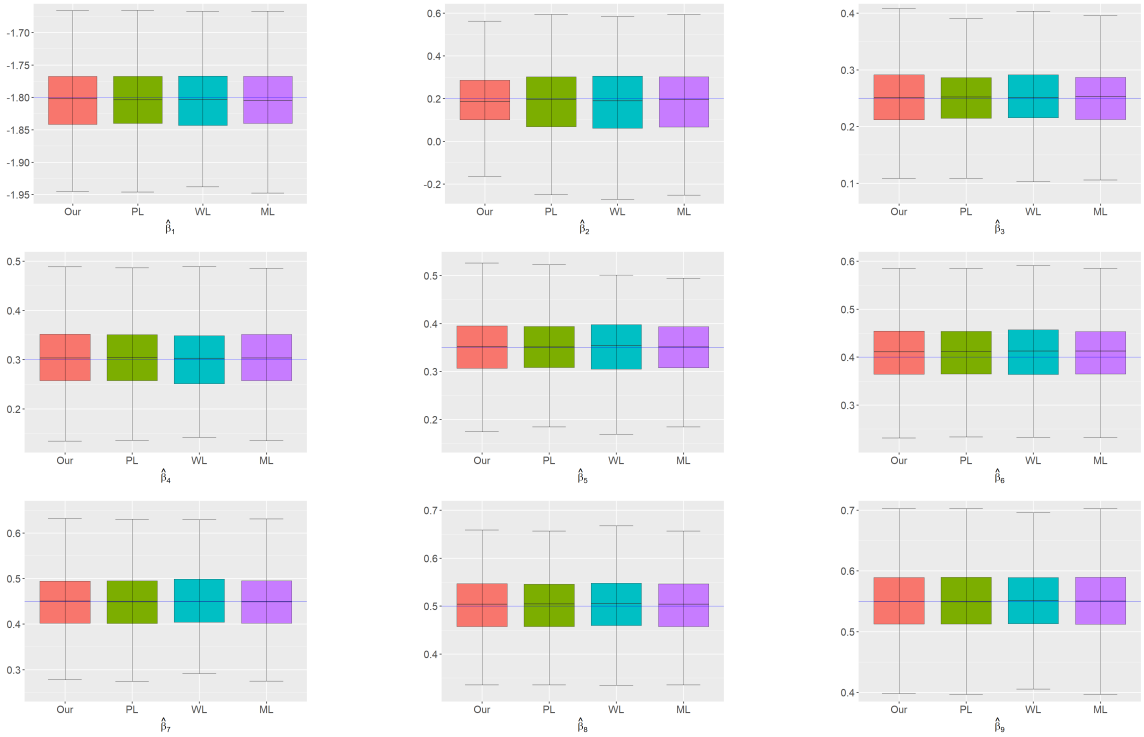

Figure 3.3: Boxplots of $\hat{\beta}_1$ to $\hat{\beta}_9$ using the four methods for the Case Control Design Sampling phase II data in Simulation 3. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. The blue horizontal line in each boxplot is the true parameter value.

Table 3.3: Empirical bias, empirical standard deviation of $\hat{\boldsymbol{\beta}}$ from four methods using the Case Control Design Sampling phase II data in Simulation 3. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. All results are timed by 100 and rounded to 3 digits.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Ours	Bias	0.34	0.83	0.22	0.45	0.07	0.96	0.02	0.23	0.16
	SD	5.31	14.52	6.05	6.82	6.70	6.83	6.82	6.19	5.36
	ESD	5.39	14.36	5.89	6.61	6.62	6.64	6.68	6.03	5.28
	CP	95%	95%	94%	94%	95%	94%	94%	94%	95%
PL	Bias	0.32	1.06	0.15	0.44	0.08	0.98	0.01	0.22	0.16
	SD	5.44	17.19	5.35	6.81	6.69	6.82	6.84	6.18	5.35
WL	Bias	0.42	1.07	0.34	0.28	0.01	1.06	0.06	0.38	0.21
	SD	5.61	18.03	5.68	7.07	6.96	7.19	6.94	6.43	5.48
ML	Bias	0.31	1.06	0.16	0.45	0.08	0.98	0.01	0.22	0.16
	SD	5.45	17.19	5.34	6.81	6.69	6.82	6.83	6.18	5.35

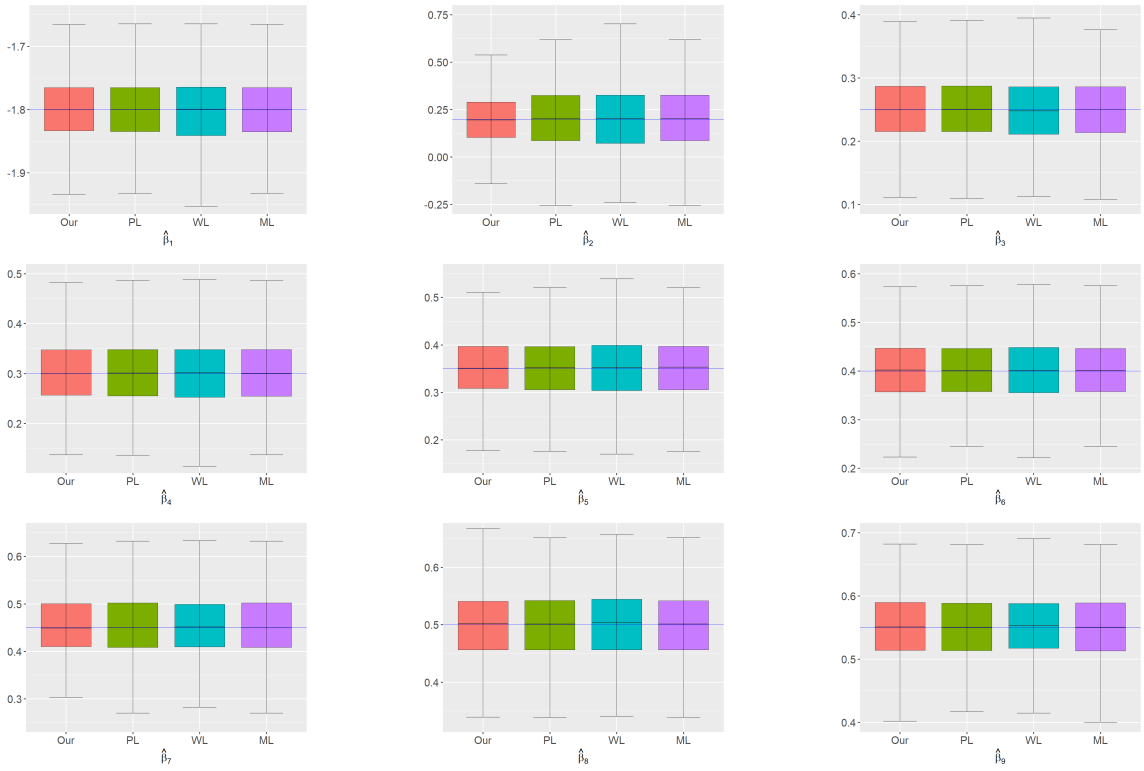

Figure 3.4: Boxplots of $\hat{\beta}_1$ to $\hat{\beta}_9$ using the four methods for the Goodness-of-Fit Based Design Sampling phase II data in Simulation 4. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. The blue horizontal line in each boxplot is the true parameter value.

Table 3.4: Empirical bias, empirical standard deviation of $\hat{\boldsymbol{\beta}}$ from four methods using the Goodness-of-Fit Based Design Sampling phase II data in Simulation 4. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. All results are timed by 100 and rounded to 3 digits.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Ours	Bias	0.16	0.05	0.09	0.14	0.17	0.40	0.5	0.08	0.18
	SD	5.26	14.92	5.37	6.61	6.27	6.39	6.68	5.99	5.38
	ESD	6.12	14.47	5.54	7.08	7.28	7.48	7.69	7.34	5.29
	CP	97%	94%	96%	97%	98%	97%	98%	98%	95%
PL	Bias	0.13	0.64	0.11	0.16	0.19	0.42	0.53	0.06	0.18
	SD	5.34	17.25	5.41	6.71	6.33	6.53	6.75	6.19	5.38
WL	Bias	0.20	0.70	0.21	0.22	0.27	0.38	0.51	0.09	0.37
	SD	5.49	18.08	5.64	6.99	6.63	6.68	7.03	6.42	5.54
ML	Bias	0.13	0.65	0.14	0.16	0.20	0.41	0.53	0.06	0.18
	SD	5.34	17.26	5.41	6.71	6.34	6.53	6.75	6.19	5.38

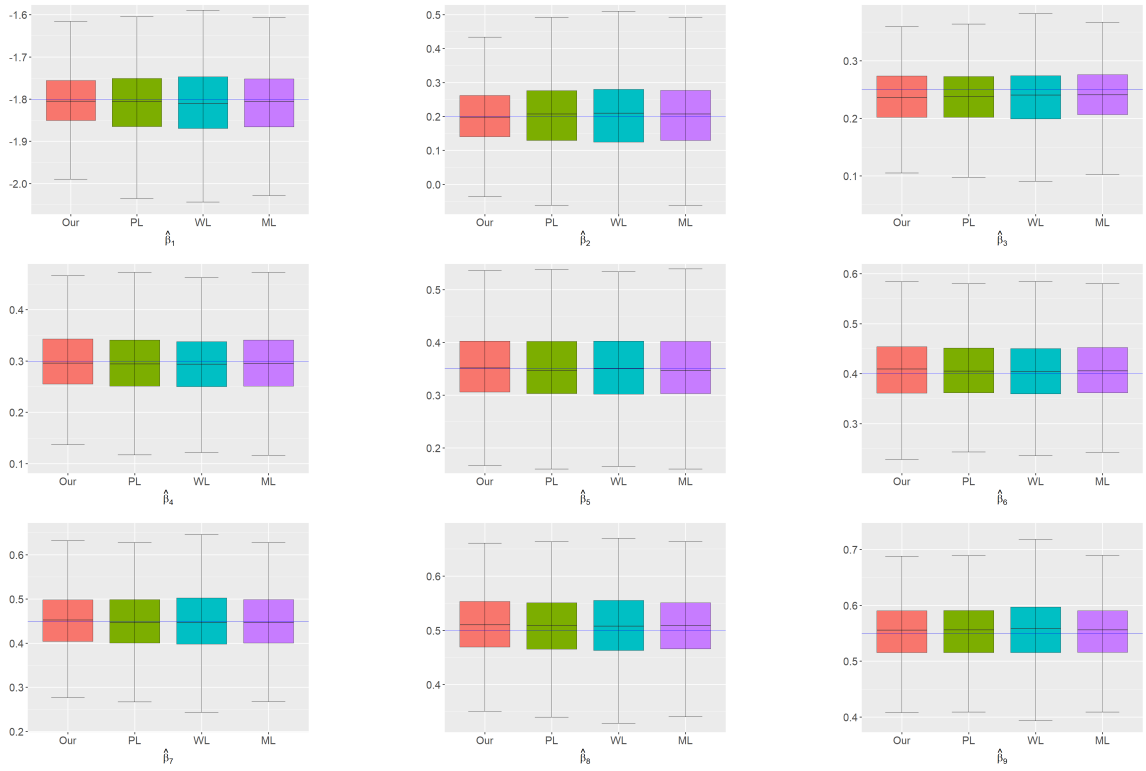

Figure 3.5: Boxplots of $\hat{\beta}_1$ to $\hat{\beta}_9$ using the four methods for the Balanced Goodness-of-Fit Based Design Sampling phase II data in Simulation 5. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. The blue horizontal line in each boxplot is the true parameter value.

Table 3.5: Empirical bias, empirical standard deviation of $\hat{\boldsymbol{\beta}}$ from four methods using the Balanced Goodness-of-Fit Based Design Sampling phase II data in Simulation 5. X is from multivariate normal distribution. X and Z are dependent. We use 4 strata for all methods. All results are timed by 100 and rounded to 3 digits.

		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
Ours	Bias	0.62	0.43	1.44	0.04	0.50	0.77	0.18	1.06	0.44
	SD	7.59	9.51	5.11	6.71	6.99	6.71	7.00	6.31	5.67
	ESD	7.41	9.31	5.77	7.31	7.50	7.67	7.84	7.48	5.55
	CP	94%	94%	97%	97%	97%	98%	97%	98%	93%
PL	Bias	0.82	0.49	1.31	0.13	0.32	0.54	0.07	0.78	0.45
	SD	8.16	11.00	5.12	6.77	7.12	6.79	7.18	6.39	5.67
WL	Bias	0.92	0.29	1.24	0.29	0.45	0.56	0.04	0.82	0.54
	SD	8.44	11.56	5.48	7.15	7.41	7.07	7.61	6.58	6.01
ML	Bias	0.90	0.49	0.96	0.13	0.33	0.54	0.07	0.77	0.45
	SD	8.14	10.00	5.09	6.77	7.12	6.80	7.18	6.39	5.67

Bibliography

Agresti, A., Ohman, P. & Caffo, B. (2004), 'Examples in which misspecification of a random effects distribution reduces efficiency', *Computational Statistics & Data Analysis* 47, 639–653.

Ando, T. & Lin, K.-C. (2014), 'A model-averaging approach for high-dimensional regression', *Journal of the American Statistical Association* 109, 254–265.

Bates, D., Maechler, M. & Bolker, B. (2016), 'lme4: Linear mixed-effects models using 'eigen' and s4', R package version 1.1-12, https://cran.r-project.org/package=lme4.

Battey, H., Fan, J., Liu, H., Lu, J. & Zhu, Z. (2018), 'Distributed testing and estimation under sparse high dimensional models', *Annals of statistics* 46(3), 1352.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* 24, 123–140.

Breslow, N. E. & Cain, K. C. (1988), 'Logistic regression for two-stage case-control data', *Biometrika* 75, 11–20.

Breslow, N. E. & Holubkov, R. (1997), 'Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling', *Journal of the Royal Statistical Society. Series B* 59, 447–461.

Chen, J., Li, D., Linton, O. & Lu, Z. (2018), 'Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series', *Journal of the American Statistical Association* 113(522), 919–932.

Claeskens, G. & Hjort, N. L. (2008), *Model Selection and Model Averaging*, Cambridge University Press, Cambridge, United Kingdom.

Downing, K., Chan, S., Downing, W., Kwong, T. & Lam, T. (2008), 'Measuring gender differences in cognitive functioning', *Multicultural Education & Technology Journal* 2, 4–18.

Garcia, T. P. & Ma, Y. (2016), 'Optimal estimator for logistic model with distribution-free random intercept', *Scandinavian Journal of Statistics* 43, 156–171.

Gerda Claeskens, Jan Magnus, A. V. & Wang, W. (2016), 'The forecast combination puzzle: A simple theoretical explanation', *International Journal of Forecasting* 32, 754–762.

Group, H. S. (1996), 'Unified huntington's disease rating scale: Reliability and Consistency', *Movement Disorders* 11, 136–142.

Hansen, B. E. (2007), 'Least squares model averaging', *Econometrica* 75(4), 1175–1189.

Heagerty, P. J. & Kurland, B. F. (2001), 'Misspecified maximum likelihood estimates and generalised linear mixed models', *Biometrika* 88, 973–985.

Hjort, N. L. & Claeskens, G. (2003), 'Frequentist model average estimators', *Journal of the American Statistical Association* 98(464), 879–899.

H.Wolpert, D. (1992), 'Stacked generalization', *Neural Networks* 5, 241–259.

JE., W. (1982), 'A two stage design for the study of the relationship between a rare exposure and a rare disease', *American Journal of Epidemiology* 115, 119–128.

JS, P., MM, S., JD, L., investigators, P. H. & of the Huntington Study Group, C. (2013), 'Cognitive decline in prodromal huntington disease: Implications for clinical trials', *Journal of Neurology, Neurosurgery, and Psychiatry* Nov, 1233–1239.

Li, R., Lin, D. K. & Li, B. (2013), 'Statistical inference in massive data sets', *Applied Stochastic Models in Business and Industry* 29(5), 399–409.

Litiére, S., Alonso, A. & Molenberghs, G. (2007), 'Type I and Type II error under random-effects misspecification in generalized linear mixed models', *Biometrics* 63, 1038–1044.

Litiére, S., Alonso, A. & Molenberghs, G. (2008), 'The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models', *Statistics in Medicine* 27, 3125–3144.

Liu, C.-A. (2015), 'Distribution theory of the least squares averaging estimator', *Journal of Econometrics* 186(1), 142–159.

Liu, Q., Yao, Q. & Zhao, G. (2020), 'Model averaging estimation for conditional volatility models with an application to stock market volatility forecast', *Journal of Forecasting* 39, 841–863.

Mitra, P., Lian, H., Mitra, R., Liang, H. & Xie, M. (2019), 'A general framework for freqentist model averaging', *SCIENCE CHINA Mathematics* 62(4), 205–226.

Neuhaus, J. M., Hauck, W. W. & Kalbfleisch, J. D. (1992), 'The effects of mixture distribution misspecification when fitting mixed-effects logistic models', *Biometrika* 79, 755–762.

Neuhaus, J. M., McCulloch, C. E. & Boylan, R. (2011), 'A note on type ii error under random effects misspecification in generalized linear mixed models', *Biometrics* 67, 654–660.

Neyman, J. (1938), 'Contribution to the theory of sampling human populations', *Journal of the American Statistical Association* 33, 101–116.

Roos, R. A. (2010), 'Huntington's disease: a clinical review', *Orphanet Journal of Rare Diseases* 5:40.

S. T. Buckland, K. P. B. & Augustin, N. H. (1997), 'Model selection: An integral part of inference', *Biometrics* 53, 603–618.

Scott, A. J. & Wild, C. J. (1997), 'Fitting regression models to case-control data by maximum likelihood', *Biometrika* 84, 57–71.

Tao, R., Zeng, D. & Lin, D. Y. (2017), 'Efficient semiparametric inference under two-phase sampling, with applications to genetic association studies', *Journal of the American Statistical Association* 112, 1468–1476.

Tsiatis, A. (2006), *Semiparametric Theory and Missing Data*, Springer, New York City, New York.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* 50(1), 1–25.

Yang, Y. (2001), 'Adaptive regression by mixing', *Journal of the American Statistical Association* 96(454), 574–588.

Zhang, D. & Davidian, M. (2001), 'Linear mixed models with flexible distributions of random effects for longitudinal data', *Biometrics* 57, 795–802.

Zhang, D. & Xia, Z. (2019), 'Weighted-averaging estimator for possible threshold in segmented linear regression model', *Journal of Statistical Planning and Inference* 200, 102–118.

Zhang, P., PX, S., A, Q. & T., G. (2008), 'Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models', *Biometrics* 64, 29–38.

Zhang, X. (2010), 'Model averaging and its applications', *Ph.D. Thesis. Academy of Mathematics and Systems Science, Chinese Academy of Sciences* .

Zhang, X., Chiou, J.-M. & Ma, Y. (2018), 'Functional prediction through averaging estimated functional linear regression models', *Biometrika* 105(4), 945–962.

Appendix A

Proof in Chapter 1

A.1 Proof of Theorem 1

Because $E\{\mathbf{S}^*_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}_0)\}=\mathbf{0}$, we can expand around $\boldsymbol{\beta}_0$ to obtain

$$\begin{aligned}
\mathbf{0} &= n^{-1/2}\sum_{i=1}^{n}\mathbf{S}^*_{eff}(\mathbf{y}_i,\mathbf{x}_i,\mathbf{z}_i,\widehat{\boldsymbol{\beta}}) \\
&= n^{-1/2}\sum_{i=1}^{n}\mathbf{S}^*_{eff}(\mathbf{y}_i,\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\beta}_0)+n^{-1}\sum_{i=1}^{n}\frac{\partial\mathbf{S}^*_{eff}(\mathbf{y}_i,\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
&= n^{-1/2}\sum_{i=1}^{n}\mathbf{S}^*_{eff}(\mathbf{y}_i,\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\beta}_0)+E\left\{\frac{\partial\mathbf{S}^*_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right\}n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+o_p(1),
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies on the line connecting $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Therefore,

$$n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)=E\left\{\frac{\partial\mathbf{S}^*_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}^{\mathrm{T}}}\right\}^{-1}n^{-1/2}\sum_{i=1}^{n}\mathbf{S}^*_{eff}(\mathbf{y}_i,\mathbf{x}_i,\mathbf{z}_i,\boldsymbol{\beta}_0)+o_p(1).$$

This implies that $n^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\rightarrow\text{Normal}(0,\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-T})$,$\mathbf{A}=E\{\partial\mathbf{S}^*_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}^{\mathrm{T}}\}$ and $\mathbf{B}=\text{var}\{\mathbf{S}^*_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}_0)\}$. Finally, it is easy to check that when $f^*_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})=f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})$, $-\mathbf{A}=-E\{\partial\mathbf{S}_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}^{\mathrm{T}}\}=\text{var}\{\mathbf{S}_{eff}(\mathbf{Y},\mathbf{X},\mathbf{Z},\boldsymbol{\beta}_0)\}=\mathbf{B}$ since $\mathbf{S}_{eff}$ is the efficient score vector. Hence, the variance-covariance simplifies to $\mathbf{B}^{-1}$ and the estimator is efficient. □

A.2 Proof of Theorem 2

Sufficiency:

$$
\begin{aligned}
& f_{\mathbf{U}|\mathbf{W},\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{w}, \mathbf{r}, \mathbf{x}, \mathbf{z}) \\
= \; & \mathrm{pr}(\mathbf{U} = \mathbf{u} \mid \mathbf{W} = \mathbf{w}, \mathbf{R} = \mathbf{r}, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) \\
= \; & \frac{\exp\{\sum_{j=1}^{m} Y_{ij}\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{r}^{\mathrm{T}}(\sum_{j=1}^{m} Y_{ij}\mathbf{Z}_{ij})\}/[\prod_{j=1}^{m}\{1+\exp(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{r})\}]\Big|_{\mathbf{Z}_i\mathbf{Y}_i=(\mathbf{w}^{\mathrm{T}},\mathbf{u}^{\mathrm{T}})^{\mathrm{T}}}}{\sum_{\mathbf{Y}_i s.t. \mathbf{Z}_i\mathbf{Y}_i=\mathbf{w}} \exp\{\sum_{j=1}^{m} Y_{ij}\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{r}^{\mathrm{T}}(\sum_{j=1}^{m} Y_{ij}\mathbf{Z}_{ij})\}/[\prod_{j=1}^{m}\{1+\exp(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{r})\}]} \\
= \; & \frac{\exp\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})(\mathbf{Z}_{iL}\mathbf{w} - \mathbf{Z}_{iL}^{-1}\mathbf{Z}_{iR}\mathbf{u}) + (\mathbf{X}_{i(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{im}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u} + \mathbf{r}^{\mathrm{T}}\mathbf{w}\}}{\sum_{\mathbf{u}} \exp\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})(\mathbf{Z}_{iL}\mathbf{w} - \mathbf{Z}_{iL}^{-1}\mathbf{Z}_{iR}\mathbf{u}) + (\mathbf{X}_{i(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{im}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u} + \mathbf{r}^{\mathrm{T}}\mathbf{w}\}} \\
= \; & \frac{\exp\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})(-\mathbf{Z}_{iL}^{-1}\mathbf{Z}_{iR}\mathbf{u}) + (\mathbf{X}_{i(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{im}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u}\}}{\sum_{\mathbf{u}} \exp\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})(-\mathbf{Z}_{iL}^{-1}\mathbf{Z}_{iR}\mathbf{u}) + (\mathbf{X}_{i(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{im}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u}\}} \\
= \; & f_{\mathbf{U}|\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{x}, \mathbf{z}).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& f_{\mathbf{R}|\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{u}, \mathbf{w}, \mathbf{x}, \mathbf{z}) \\
= \;& \mathrm{pr}(\mathbf{R} = \mathbf{r} \mid \mathbf{U} = \mathbf{u}, \mathbf{W} = \mathbf{w}, \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) \\
= \;& \Big(f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}) \exp\{\sum_{j=1}^{m} Y_{ij} \mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{r}^{\mathrm{T}} (\sum_{j=1}^{m} Y_{ij} \mathbf{Z}_{ij})\} \\
& / [\prod_{j=1}^{m} \{1 + \exp(\mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}} \mathbf{r})\}] \Big|_{\mathbf{Z}_i \mathbf{Y}_i = (\mathbf{w}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}} \Big) \\
& / \Big(\int f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}) \exp\{\sum_{j=1}^{m} Y_{ij} \mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{r}^{\mathrm{T}} (\sum_{j=1}^{m} Y_{ij} \mathbf{Z}_{ij})\} \\
& / [\prod_{j=1}^{m} \{1 + \exp(\mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}} \mathbf{r})\}] \Big|_{\mathbf{Z}_i \mathbf{Y}_i = (\mathbf{w}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}} d\mathbf{r} \Big) \\
= \;& \frac{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}) \exp(\mathbf{w}^{\mathrm{T}} \mathbf{r}) / [\prod_{j=1}^{m} \{1 + \exp(\mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}} \mathbf{r})\}]}{\int f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}) \exp(\mathbf{w}^{\mathrm{T}} \mathbf{r}) / [\prod_{j=1}^{m} \{1 + \exp(\mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}} \mathbf{r})\}] d\mathbf{r}} \\
= \;& f_{\mathbf{R}|\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{w}, \mathbf{x}, \mathbf{z}).
\end{aligned}
$$

Completeness:

$$
\begin{aligned}
& E\{\mathbf{a}(\mathbf{W},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} \\
= & \int \mathbf{a}(\mathbf{w},\mathbf{X},\mathbf{Z}) f_{\mathbf{W}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{w} \mid \mathbf{R},\mathbf{X},\mathbf{Z}) d\mu(\mathbf{w}) \\
= & \int \mathbf{a}(\mathbf{w},\mathbf{X},\mathbf{Z}) \frac{\exp\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\cdot,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})(\mathbf{Z}_{iL}\mathbf{w} - \mathbf{Z}_{iL}^{-1}\mathbf{Z}_{iR}\mathbf{u}) + (\mathbf{X}_{i(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\cdot,\mathbf{X}_{im}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u} + \mathbf{R}^{\mathrm{T}}\mathbf{w}\}}{\prod_{j=1}^{m}\{1+\exp(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{R})\}} \\
& d\mu(\mathbf{u}) d\mu(\mathbf{w}) \\
= & b(\mathbf{R},\mathbf{X}_i,\mathbf{Z}_i) \int \mathbf{a}(\mathbf{w},\mathbf{X},\mathbf{Z}) \exp[\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{Z}_{iL} + \mathbf{R}^{\mathrm{T}}\}\mathbf{w}] d\mu(\mathbf{w}),
\end{aligned}
$$

where

$$
b(\mathbf{R},\mathbf{X}_i,\mathbf{Z}_i) = \frac{\int \exp\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})(-\mathbf{Z}_{iL}^{-1}\mathbf{Z}_{iR}\mathbf{u}) + (\mathbf{X}_{i(q+1)}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{im}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{u}\} d\mu(\mathbf{u})}{\prod_{j=1}^{m}\{1+\exp(\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{R})\}}
$$

is a positive function. Thus, $E\{\mathbf{a}(\mathbf{W},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} = \mathbf{0}$ implies

$$
\int \mathbf{a}(\mathbf{W},\mathbf{X},\mathbf{Z}) \exp[\{(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta},\ldots,\mathbf{X}_{iq}^{\mathrm{T}}\boldsymbol{\beta})\mathbf{Z}_{iL} + \mathbf{R}^{\mathrm{T}}\}\mathbf{w}] d\mu(\mathbf{w}) = \mathbf{0}.
$$

This implies the Laplace transformation of $\mathbf{a}$ is zero, hence $\mathbf{a}(\mathbf{W},\mathbf{X},\mathbf{Z}) = \mathbf{0}$. □

A.3 Derivation of the Nuisance Tangent Space

Let $\mathbf{Y}_i = (Y_{i1},\ldots,Y_{im})^{\mathrm{T}}$, $\mathbf{X}_i = (\mathbf{X}_{i1},\ldots,\mathbf{X}_{im})$ denote a $p \times m$ matrix, and $\mathbf{Z}_i = (\mathbf{Z}_{i1},\ldots,\mathbf{Z}_{im})$ denote a $q \times m$ matrix. Without loss of generality, assume that the first q columns of $\mathbf{Z}_i$ form an invertable matrix. Using f to denote various densities

described by the subindices, the likelihood for the ith cluster is

$$\begin{aligned} & f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \\ = & \int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i \mid \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) d\mu(\mathbf{r}_i) \end{aligned} \tag{A.1}$$

Here, $\mu(\cdot)$ denotes the dominating measure. We want to leave $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i)$ unspecified.

We assume throughout that $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z})$ in (A.1) is an unknown density. With $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}$ as the unknown (nuisance) distribution, the nuisance tangent space is

$$\Lambda = [E\{\mathbf{h}(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\} : E(\mathbf{h}) = \mathbf{0}, E(\mathbf{h}^{\mathrm{T}}\mathbf{h}) < \infty], \tag{A.2}$$

and its orthogonal complement is

$$\Lambda^{\perp} = [\boldsymbol{\gamma}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) : E\{\boldsymbol{\gamma}(\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{R}, \mathbf{X}, \mathbf{Z}\} = \mathbf{0}, E(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\gamma})\} < \infty], \tag{A.3}$$

where $\mathbf{h}$ and $\boldsymbol{\gamma}$ are p-dimensional vectors.

To see this, we first derive the nuisance score vector of the parametric submodel of (A.1) with respect to $\boldsymbol{\eta}$:

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\eta}} \log \int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i \mid \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) d\mu(\mathbf{r}_i) \\
= & \frac{1}{\int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i \mid \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) d\mu(\mathbf{r}_i)} \times \\
& \frac{\partial}{\partial \boldsymbol{\eta}} \int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i \mid \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) d\mu(\mathbf{r}_i) \\
= & \frac{1}{f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta})} \frac{\partial}{\partial \boldsymbol{\eta}} \int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i \mid \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) d\mu(\mathbf{r}_i) \\
= & \int \frac{f_{\mathbf{Y},\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i, \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta})}{f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta})} \frac{1}{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta})} \frac{\partial}{\partial \boldsymbol{\eta}} f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) d\mu(\mathbf{r}_i) \\
= & \int \frac{\partial}{\partial \boldsymbol{\eta}} \log f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) f_{\mathbf{R}|\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) d\mu(\mathbf{r}_i) \\
= & E\{\frac{\partial}{\partial \boldsymbol{\eta}} \log f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{R}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\eta}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\} \\
= & E\{\mathbf{h}(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\}
\end{aligned}
$$

We also have

$$
\begin{aligned}
& E\{\frac{\partial}{\partial \boldsymbol{\eta}} \log f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{R}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\eta})\} \\
= & \int\int\int \frac{\partial}{\partial \boldsymbol{\eta}} \log f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) d\mu(\mathbf{r}, \mathbf{x}, \mathbf{z}) \\
= & \int\int\int \frac{1}{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta})} f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) \frac{\partial}{\partial \boldsymbol{\eta}} f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) d\mu(\mathbf{r}, \mathbf{x}, \mathbf{z}) \\
= & \frac{\partial}{\partial \boldsymbol{\eta}} \int\int\int f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}, \mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) d\mu(\mathbf{r}, \mathbf{x}, \mathbf{z}) = \frac{\partial}{\partial \boldsymbol{\eta}} 1 = 0
\end{aligned}
$$

This completes the nuisance tangent space derivation for a parametric submodel. Since the definition of the nuisance tangent space of our original model is the mean square closure of the nuisance tangent space of all parametric submodels, the desired nuisance tangent space is

$$\Lambda \;=\; [E\{\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\} : E(\mathbf{h}) = \mathbf{0}, E(\mathbf{h}^{\mathrm{T}}\mathbf{h}) < \infty],$$

Now we need to prove that for any bounded random functions $E\{\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\} \in \Lambda$, they are the nuisance score vectors of a specific parametric submodel. Suppose the true model for the unknown density $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})$ is $f_0(\mathbf{r},\mathbf{x},\mathbf{z})$, and let

$$f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) = f_0(\mathbf{r},\mathbf{x},\mathbf{z})\{1 + \boldsymbol{\eta}^{\mathrm{T}} h(\mathbf{R},\mathbf{X},\mathbf{Z})\}$$

where $\boldsymbol{\eta}$ is small enough such that

$$1 + \boldsymbol{\eta}^{\mathrm{T}} h(\mathbf{R},\mathbf{X},\mathbf{Z}) \geq 0$$

Then we have

$$\begin{aligned} & \int\int\int f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{x},\mathbf{z}) \\ = \;& \int\int\int f_0(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{x},\mathbf{z}) + \int\int\int f_0(\mathbf{r},\mathbf{x},\mathbf{z}) \boldsymbol{\eta}^{\mathrm{T}} h(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{x},\mathbf{z}) \\ = \;& 1 + \boldsymbol{\eta}^{\mathrm{T}} E\{h(\mathbf{R},\mathbf{X},\mathbf{Z})\} = 1 \end{aligned}$$

This means $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})$ is a valid probability density function as it is always positive and its integration from negative infinity to positive infinity is 1. When $\boldsymbol{\eta} = 0$, $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) = f_0(\mathbf{r},\mathbf{x},\mathbf{z})$, so the true model is contained in $f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})$. By definition,

it is a parametric submodel. Since the nuisance score vectors for the parametric submodel is

$$\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\eta}} \log \int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y}_i \mid \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) d\mu(\mathbf{r}_i) \\
= & \int \frac{\partial}{\partial \boldsymbol{\eta}} \log f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\eta}) f_{\mathbf{R}|\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{r}_i \mid \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}) d\mu(\mathbf{r}_i) \\
= & E\{\frac{\partial}{\partial \boldsymbol{\eta}} \log f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{R}, \mathbf{X}, \mathbf{Z}; \boldsymbol{\eta}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\} \\
= & E\{\mathbf{h}(\mathbf{R}, \mathbf{X}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}\}
\end{aligned}$$

We have shown that any element in the set defined in (A.2) is indeed one element in the nuisance tangent space of model (A.1).

Now we need to derive $\Lambda^{\perp}$. For any element $\boldsymbol{\gamma}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \in \Lambda^{\perp}$, we must have

$$
\begin{aligned}
0 &= E[\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{Y},\mathbf{X},\mathbf{Z})E\{\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\}] \\
&= \int\int\int \boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{y},\mathbf{x},\mathbf{z})E\{\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\} f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{x},\mathbf{z}) d\mu(\mathbf{y},\mathbf{x},\mathbf{z}) \\
&= \int\int\int \boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{y},\mathbf{x},\mathbf{z}) \int \mathbf{h}(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R}|\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{y},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{x},\mathbf{z}) d\mu(\mathbf{y},\mathbf{x},\mathbf{z}) \\
&= \int\int\int\int \boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{y},\mathbf{x},\mathbf{z})\mathbf{h}(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R}|\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{y},\mathbf{x},\mathbf{z}) f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) d\mu(\mathbf{y},\mathbf{x},\mathbf{z}) \\
&= \int\int\int\int \boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{y},\mathbf{x},\mathbf{z})\mathbf{h}(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{y},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{y},\mathbf{x},\mathbf{z}) \\
&= \int\int\int \left[\int \boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{y},\mathbf{x},\mathbf{z})\mathbf{h}(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{y})\right] f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{x},\mathbf{z}) \\
&= \int\int\int E\{\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{Y},\mathbf{X},\mathbf{Z})\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{x},\mathbf{z}) \\
&= E[E\{\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{Y},\mathbf{X},\mathbf{Z})\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\}] \\
&= \int\int\int \left[\int \boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{y},\mathbf{x},\mathbf{z}) f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{y})\right] \mathbf{h}(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{x},\mathbf{z}) \\
&= \int\int\int E\{\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\}\mathbf{h}(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r},\mathbf{x},\mathbf{z}) \\
&= E[E\{\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\}\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z})]
\end{aligned}
$$

for any $\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z})$ where $E(\mathbf{h}(\mathbf{R},\mathbf{X},\mathbf{Z})) = 0$. Thus we must have $E\{\boldsymbol{\gamma}^{\mathrm{T}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} = 0$ and $E(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\gamma}) < \infty$. Then we have

$$
\Lambda^{\perp} = [\boldsymbol{\gamma}(\mathbf{Y},\mathbf{X},\mathbf{Z}) : E\{\boldsymbol{\gamma}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} = \mathbf{0}, E(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\gamma})\} < \infty],
$$

The score function is

$$
\begin{aligned}
& \mathbf{S}_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z};\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log\{f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{x},\mathbf{z};\boldsymbol{\beta})\} \\
= & \frac{\partial}{\partial \boldsymbol{\beta}} \log \int f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & \frac{1}{f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{x},\mathbf{z})} \int \frac{\partial}{\partial \boldsymbol{\beta}} f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & \frac{1}{f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{x},\mathbf{z})} \int \{\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z})\} f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & \int \{\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z})\} \frac{f_{\mathbf{Y},\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{r},\mathbf{x},\mathbf{z})}{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{y},\mathbf{x},\mathbf{z})} f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & \int \{\frac{\partial}{\partial \boldsymbol{\beta}} \log f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{y} \mid \mathbf{r},\mathbf{x},\mathbf{z})\} f_{\mathbf{R}|\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{y},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \log\{f_{\mathbf{Y}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{Y} \mid \mathbf{R},\mathbf{X},\mathbf{Z};\boldsymbol{\beta})\} \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\right]
\end{aligned}
$$

A.1 Proof of $E\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\}$=$E\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\}$

Here we need to use the first equation of the first part of Theorem 2.

$$
\begin{aligned}
& E\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\} \\
= & \int \mathbf{h}^*(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R}|\mathbf{Y},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{y},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & \int \mathbf{h}^*(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R}|\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & \int \mathbf{h}^*(\mathbf{r},\mathbf{x},\mathbf{z}) \frac{f_{\mathbf{R},\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})} d\mu(\mathbf{r}) \\
= & \int \mathbf{h}^*(\mathbf{r},\mathbf{x},\mathbf{z}) \frac{f_{\mathbf{U}|\mathbf{R},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{r},\mathbf{w},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})} d\mu(\mathbf{r}) \\
= & \int \mathbf{h}^*(\mathbf{r},\mathbf{x},\mathbf{z}) \frac{f_{\mathbf{U}|\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{w},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})} d\mu(\mathbf{r}) \\
= & \int \mathbf{h}^*(\mathbf{r},\mathbf{x},\mathbf{z}) \frac{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z}) f_{\mathbf{R},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z}) f_{\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{w},\mathbf{x},\mathbf{z})} d\mu(\mathbf{r}) \\
= & \int \mathbf{h}^*(\mathbf{r},\mathbf{x},\mathbf{z}) f_{\mathbf{R}|\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{w},\mathbf{x},\mathbf{z}) d\mu(\mathbf{r}) \\
= & E\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\}
\end{aligned}
$$

Now we need to show why $E\{E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} \mid \mathbf{R},\mathbf{X},\mathbf{Z}\}$=$E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\}$. Here we need to use the second equation of the first part of Theorem 2.

$$
\begin{aligned}
& E\{E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} \\
= & \int E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} f_{\mathbf{W}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{w} \mid \mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{w}) \\
= & \int\int \mathbf{S}^*_\beta(\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}) f_{\mathbf{U}|\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u} \mid \mathbf{w},\mathbf{x},\mathbf{z}) d\mu(\mathbf{u}) f_{\mathbf{W}|\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{w} \mid \mathbf{r},\mathbf{x},\mathbf{z}) d\mu(\mathbf{w}) \\
= & \int\int \mathbf{S}^*_\beta(\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}) \frac{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{w},\mathbf{x},\mathbf{z})} \frac{f_{\mathbf{W},\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{w},\mathbf{r},\mathbf{x},\mathbf{z})}{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})} d\mu(\mathbf{u}) d\mu(\mathbf{w}) \\
= & \int\int \mathbf{S}^*_\beta(\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}) f_{\mathbf{R}|\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{w},\mathbf{x},\mathbf{z}) \frac{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})} d\mu(\mathbf{u}) d\mu(\mathbf{w}) \\
= & \int\int \mathbf{S}^*_\beta(\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}) f_{\mathbf{R}|\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r} \mid \mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z}) \frac{f_{\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})} d\mu(\mathbf{u}) d\mu(\mathbf{w}) \\
= & \int\int \mathbf{S}^*_\beta(\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}) \frac{f_{\mathbf{R},\mathbf{U},\mathbf{W},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{u},\mathbf{w},\mathbf{x},\mathbf{z})}{f_{\mathbf{R},\mathbf{X},\mathbf{Z}}(\mathbf{r},\mathbf{x},\mathbf{z})} d\mu(\mathbf{u}) d\mu(\mathbf{w}) \\
= & E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\}
\end{aligned}
$$

Since we constructed $\mathbf{h}^*$ such that

$$E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} = E[E^*\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\} \mid \mathbf{R},\mathbf{X},\mathbf{Z}].$$

We have

$$E\{E\{\mathbf{S}^*_\beta(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} \mid \mathbf{R},\mathbf{X},\mathbf{Z}\} - E[E^*\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{Y},\mathbf{X},\mathbf{Z}\} \mid \mathbf{R},\mathbf{X},\mathbf{Z}] = 0.$$

which implies that

$$E[E\{\mathbf{S}^*_{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} - E^*\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} \mid \mathbf{R},\mathbf{X},\mathbf{Z}] = 0.$$

Now we use the second part of Theorem 2, we have

$$E\{\mathbf{S}^*_{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} - E^*\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} = 0.$$

This means the conditional expectation of $h^*(\mathbf{R},\mathbf{X},\mathbf{Z})$ given $(\mathbf{W},\mathbf{X},\mathbf{Z})$ satisfies

$$E^*\{\mathbf{h}^*(\mathbf{R},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\} \;=\; E\{\mathbf{S}^*_{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X},\mathbf{Z}) \mid \mathbf{W},\mathbf{X},\mathbf{Z}\}$$

Appendix B

Proof in Chapter 2

B.1 Proof of asymptotic unbiasedness of $CV(\mathbf{w})$ in estimating $\mathrm{E}[\{\widehat{Y}(\mathbf{w}) - Y\}^2]$

First of all, we have the familiar decompositions

$$\begin{aligned}
\mathrm{E}\{CV(\mathbf{w})\} &= \mathrm{E}[\{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2] \\
&= \mathrm{E}[\{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) - Y_{1i} + \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2] \\
&= \mathrm{E}[\{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2] + \mathrm{E}[\{Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2]
\end{aligned}$$

and

$$\mathrm{E}[\{\widehat{Y}(\mathbf{w}) - Y\}^2] = \mathrm{E}[\{\widehat{Y}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2] + \mathrm{E}[\{Y - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2].$$

Because the observations $\{\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}, Y_{1i}\}, i = 1, \ldots, n_1$, and $\{\mathbf{X}, \mathbf{Z}^{[1]}, Y\}$ are iid, we have that $\mathrm{E}[\{Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2] = \mathrm{E}[\{Y - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2]$ and $\mathrm{E}[\{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2] = \mathrm{E}[\{\widehat{Y}^{-}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2]$, where $\widehat{Y}^{-}(\mathbf{w})$ stands for the quantity obtained in the same way as $\widehat{Y}(\mathbf{w})$ except that we use $n_1 - 1$ observations randomly drawn from the first population,

instead of the n_1 observations. When $n_1 \to \infty$, $\widehat{Y}(\mathbf{w}) - \widehat{Y}^{-}(\mathbf{w}) \to 0$ in probability, hence $\mathrm{E}\{CV(\mathbf{w})\} = \mathrm{E}[\{\widehat{Y}(\mathbf{w}) - Y\}^2] + o(1)$ for any $\mathbf{w}$. □

B.2 Proof of Theorem 3

We fist prove two preliminary results. By Assumption 2,

$$Y^{\star}_{1i,j}\{Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i)\} = g(\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i; \boldsymbol{\beta}^{\star}_j, \boldsymbol{\alpha}^{\star}_1, \boldsymbol{\gamma}^{\star}_1)\epsilon_i$$

and

$$\begin{aligned}
&\{Y^{\star}_{1i,j} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i)\}\{Y^{\star}_{1i,j'} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i)\} \\
&\quad = \left\{g(\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i; \boldsymbol{\beta}^{\star}_j, \boldsymbol{\alpha}^{\star}_1, \boldsymbol{\gamma}^{\star}_1) - g_{\mathrm{true}}(\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1)\right\} \\
&\qquad \times \left\{g(\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i; \boldsymbol{\beta}^{\star}_{j'}, \boldsymbol{\alpha}^{\star}_1, \boldsymbol{\gamma}^{\star}_1) - g_{\mathrm{true}}(\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i; \boldsymbol{\beta}, \boldsymbol{\alpha}_1, \boldsymbol{\gamma}_1)\right\},
\end{aligned}$$

we know that

$$\begin{aligned}
&\textit{the variances of } Y^{\star}_{1i,j}\{Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i)\} \textit{ and} \\
&\{Y^{\star}_{1i,j} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i)\}\{Y^{\star}_{1i,j'} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}^{[1]}_i)\} \\
&\textit{are bounded by a constant uniformly for all } j, j' \in \{1, \ldots, N\}.
\end{aligned} \tag{B.1}$$

Second, note that

$$\begin{aligned}
\left\| \frac{\partial \{\widehat{Y}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2}{\partial \widehat{\boldsymbol{\theta}}} \right\| &= 2 \sum_{j=1}^{N} w_j \{\widehat{Y}_j - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\} \sum_{j=1}^{N} w_j \left\| \frac{\partial \widehat{Y}_j}{\partial \widehat{\boldsymbol{\theta}}_{[j]}} \right\| \\
&\leq \left\{ 2 \max_j |\widehat{Y}_j| + 2|\mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})| \right\} \max_j \left\| \frac{\partial \widehat{Y}_j}{\partial \widehat{\boldsymbol{\theta}}_{[j]}} \right\|, \\
\left\| \frac{\partial \{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2}{\partial \widehat{\boldsymbol{\theta}}} \right\| &= 2 \sum_{j=1}^{N} w_j \{\widehat{Y}_{1i}^{(-i)} - Y_{1i}\} \sum_{j=1}^{N} \widehat{w}_j \left\| \frac{\partial \widehat{Y}_{1i}^{(-i)}}{\partial \widehat{\boldsymbol{\theta}}_{[j]}^{(-i)}} \right\| \\
&\leq \left\{ 2 \max_j |\widehat{Y}_{1i}^{(-i)}| + 2|\mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})| + 2|\epsilon_i| \right\} \max_j \left\| \frac{\partial \widehat{Y}_{1i}^{(-i)}}{\partial \widehat{\boldsymbol{\theta}}_{[j]}^{(-i)}} \right\|, \\
\widehat{Y}_{j,1i}^{(-i)} &= g(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \widehat{\boldsymbol{\beta}}_{[j]}, \widehat{\boldsymbol{\alpha}}_1^{(-i)}, \widehat{\boldsymbol{\gamma}}_1^{(-i)}), \quad j \geq 2, \\
\widehat{Y}_{1,1i}^{(-i)} &= g(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \widehat{\boldsymbol{\beta}}_{[1]}^{(-i)}, \widehat{\boldsymbol{\alpha}}_1^{(-i)}, \widehat{\boldsymbol{\gamma}}_1^{(-i)}), \\
\widehat{Y}_{j,1i} &= g(\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}; \widehat{\boldsymbol{\beta}}_{[j]}, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\gamma}}_1),
\end{aligned}$$

where $\widehat{\boldsymbol{\theta}}_{[j]} = (\widehat{\boldsymbol{\beta}}_{[j]}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}})^{\mathrm{T}}$ for $j = 1, \ldots, J$, $\widehat{\boldsymbol{\theta}}_{[1]}^{(-i)} = (\widehat{\boldsymbol{\beta}}_{[1]}^{(-i)^{\mathrm{T}}}, \widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_1^{\mathrm{T}})^{\mathrm{T}}$, $\widehat{\boldsymbol{\theta}}_{[j]}^{(-i)} = \widehat{\boldsymbol{\theta}}_{[j]}$ for $j = 2, \ldots, J$, $\widehat{\boldsymbol{\beta}}_{[1]}^{(-i)}$ is the estimate of $\boldsymbol{\beta}$ under the main model without using the-ith-observation. Then, from Assumptions 2 and 3, we know that

$$\begin{gathered}
\textit{the components of the derivatives } \partial \{\widehat{Y}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 / \partial \widehat{\boldsymbol{\theta}} \,|_{\widehat{\boldsymbol{\theta}} = \overline{\boldsymbol{\theta}}} \textit{ and} \\
\partial \{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2 / \partial (\widehat{\boldsymbol{\beta}}_{[1]}^{(-i)^{\mathrm{T}}}, \widehat{\boldsymbol{\beta}}_{[2]}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_{[N]}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}_1^{(-i)^{\mathrm{T}}}, \widehat{\boldsymbol{\gamma}}_1^{(-i)^{\mathrm{T}}})^{\mathrm{T}} \,|_{(\widehat{\boldsymbol{\beta}}_{[1]}^{(-i)^{\mathrm{T}}}, \widehat{\boldsymbol{\beta}}_{[2]}^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_{[N]}^{\mathrm{T}}, \widehat{\boldsymbol{\alpha}}_1^{(-i)^{\mathrm{T}}}, \widehat{\boldsymbol{\gamma}}_1^{(-i)^{\mathrm{T}}})^{\mathrm{T}} = \overline{\boldsymbol{\theta}}} \\
\textit{are } O_p(1) \textit{ uniformly for any } \overline{\boldsymbol{\theta}} \textit{ in a local neighborhood of } \boldsymbol{\theta}^\star \\
\textit{and uniformly for any } \mathbf{w} \in \mathcal{W}, \qquad \text{(B.2)}
\end{gathered}$$

where $\widehat{\boldsymbol{\alpha}}_1^{(-i)}$ and $\widehat{\boldsymbol{\gamma}}_1^{(-i)}$ are the estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ respectively under the main model without using the-ith-observation. In addition,

$$\begin{aligned}
& \mathrm{E}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left|\{\widehat{Y}(\mathbf{w})-\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}^2-\{Y^\star(\mathbf{w})-\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}^2\right|\right] \\
= \;& \mathrm{E}\left[\sup_{\mathbf{w}\in\mathcal{W}}\left|\{\widehat{Y}(\mathbf{w})-Y^\star(\mathbf{w})\}\{\widehat{Y}(\mathbf{w})+Y^\star(\mathbf{w})-2\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}\right|\right] \\
\leq \;& \mathrm{E}\left\{\sup_{1\leq j\leq N}|\widehat{Y}_j-Y_j^\star|\sup_{1\leq j\leq N}|\widehat{Y}_j+Y_j^\star-2\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})|\right\} \\
\leq \;& \mathrm{E}\left(\sup_{1\leq j\leq N}|\widehat{Y}_j-Y_j^\star|^2\right)+2\mathrm{E}\left[\left\{\sup_{1\leq j\leq N}|\widehat{Y}_j-Y_j^\star|\right\}\sup_{1\leq j\leq N}|Y_j^\star-\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})|\right] \\
\leq \;& 2\mathrm{E}\left(\sup_{1\leq j\leq N}|\widehat{Y}_j-Y_j^\star|^2\right)+\mathrm{E}\left\{\sup_{1\leq j\leq N}|Y_j^\star-\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})|^2\right\}. \qquad \text{(B.3)}
\end{aligned}$$

It is seen that

$$\begin{aligned}
& \xi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\{\widehat{Y}(\mathbf{w})-\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}^2-\{Y^\star(\mathbf{w})-\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}^2\right| \\
= \;& \xi^{-1}\sup_{\mathbf{w}\in\mathcal{W}}\left|\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star\right)^{\mathrm{T}}\frac{\partial\{\widehat{Y}(\mathbf{w})-\mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}^2}{\partial\widehat{\boldsymbol{\theta}}}\,|_{\widehat{\boldsymbol{\theta}}=\widetilde{\boldsymbol{\theta}}}\right| \\
= \;& \xi^{-1}O_p(\underline{n}^{-1/2}N^{3/2}) \\
= \;& o_p(1), \qquad \text{(B.4)}
\end{aligned}$$

where the first equality uses (B.2) and Assumption 3, the second equality uses (B.2) and Assumptions 1-3, and the third equality uses Assumption 5. Further, we have

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \frac{|R(\mathbf{w}) - R^\star(\mathbf{w})|}{R^\star(\mathbf{w})} \\
\leq\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} |R(\mathbf{w}) - R^\star(\mathbf{w})| \\
=\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \mathrm{E}\left[\{\widehat{Y}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 - \{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 \right] \right| \\
\leq\ & \mathrm{E}\left(\xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \{\widehat{Y}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 - \{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 \right| \right) \\
=\ & o(1), \qquad \text{(B.5)}
\end{aligned}$$

where the second inequality uses (B.3) and Assumption 4, and the last equality is due to (B.4).

Similarly to (B.4), we can obtain that

$$\sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2 - \{Y_{1i}^\star(\mathbf{w}) - Y_{1i}\}^2 \right] \right| = O_p(\underline{n}^{-1/2} N^{3/2}). \qquad \text{(B.6)}$$

Let

$$\begin{aligned}
g_{1i} &= \left\{ Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \left\{ Y_{1i} + \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\}, \\
CV^\star(\mathbf{w}) &= CV(\mathbf{w}) - \frac{1}{n_1} \sum_{i=1}^{n_1} g_{1i}, \qquad \text{(B.7)}
\end{aligned}$$

where $n_1^{-1} \sum_{i=1}^{n_1} g_{1i}$ is unrelated to $\mathbf{w}$, so

$$\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} CV^\star(\mathbf{w}). \qquad \text{(B.8)}$$

It is seen that

$$
\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}} \frac{|CV^\star(\mathbf{w}) - R^\star(\mathbf{w})|}{R^\star(\mathbf{w})} \\
\leq\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} |CV^\star(\mathbf{w}) - R^\star(\mathbf{w})| \\
=\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2 - g_{1i} \right] - \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 \right] \right| \\
\leq\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\{Y_{1i}^\star(\mathbf{w}) - Y_{1i}\}^2 - g_{1i} \right] - \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 \right] \right| \\
& + \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \{\widehat{Y}_{1i}^{(-i)}(\mathbf{w}) - Y_{1i}\}^2 - \{Y_{1i}^\star(\mathbf{w}) - Y_{1i}\}^2 \right\} \right| \\
=\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\{Y_{1i}^\star(\mathbf{w}) - Y_{1i}\}^2 - g_{1i} \right] - \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 \right] \right| \\
& + \xi^{-1} O_p(\underline{n}^{-1/2} N^{3/2}) \\
\leq\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{Y_{1i}^\star(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2 - \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 \right] \right| \\
& + \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\{Y_{1i}^\star(\mathbf{w}) - Y_{1i}\}^2 - g_{1i} - \{Y_{1i}^\star(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2 \right] \right| \\
& + \xi^{-1} O_p(\underline{n}^{-1/2} N^{3/2}) \\
=\ & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{Y_{1i}^\star(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2 - \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2 \right] \right| \\
& + \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[2Y_{1i}^\star(\mathbf{w}) \left\{ Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \right] \right| \\
& + \xi^{-1} O_p(\underline{n}^{-1/2} N^{3/2}),
\end{aligned}
\tag{B.9}
$$

where the second equality is from (B.6). By Lemma 1 of Zhang (2010), (B.5), (B.9) and Assumption 5, to prove (5), it suffices to show

$$\begin{aligned} & \xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{Y_{1i}^{\star}(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2 - \mathrm{E}\left[\{Y^{\star}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right] \right| \\ = \; & o_p(N^{-1/2}) = o_p(1) \end{aligned} \tag{B.10}$$

and

$$\xi^{-1} \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y_{1i}^{\star}(\mathbf{w}) \left\{ Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \right] \right| = O_p(\underline{n}^{-1/2} N) = o_p(1). \tag{B.11}$$

We consider (B.10) at first. It is seen that

$$
\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{Y_{1i}^{\star}(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2 - \mathrm{E}\left[\{Y^{\star}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right] \right| \\
= & \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{\sum_{j=1}^{N} w_j Y_{1i,j}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2 - \mathrm{E}\left[\{\sum_{j=1}^{N} w_j Y_j^{\star} - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right] \right| \\
= & \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} [\sum_{j=1}^{N} w_j \{Y_{1i,j}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}]^2 - \mathrm{E}\left([\sum_{j=1}^{N} w_j \{Y_j^{\star} - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}]^2\right) \right| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^{N} \sum_{j'=1}^{N} w_j w_{j'} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \{Y_{1i,j}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}\{Y_{1i,j'}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\} \right. \\
& \qquad \left. - \mathrm{E}\left[\{Y_j^{\star} - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}\{Y_{j'}^{\star} - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}\right] \right| \\
= & n_1^{-1/2} \sup_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^{N} \sum_{j'=1}^{N} w_j w_{j'} \widehat{\pi}_{j,j'} \\
\leq & n_1^{-1/2} \sup_{j,j'} \widehat{\pi}_{j,j'} \\
= & o_p(\xi N^{-1/2}),
\end{aligned}
\tag{B.12}
$$

where

$$
\begin{aligned}
\widehat{\pi}_{j,j'} \equiv & \left| n_1^{-1/2} \sum_{i=1}^{n_1} \{Y_{1i,j}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}\{Y_{1i,j'}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\} \right. \\
& \qquad \left. - \mathrm{E}\left[\{Y_j^{\star} - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}\{Y_{j'}^{\star} - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}\right] \right|.
\end{aligned}
$$

The last step in (B.12) is because by (B.1) and the Chebyshev's inequality, we have that for any $\delta > 0$,

$$\begin{aligned}
& \mathrm{pr}\left\{\xi^{-1}n_1^{-1/2}\sup_{j,j'}\widehat{\pi}_{j,j'} > \delta N^{-1/2}\right\} \\
\leq & \sum_{j=1}^{N}\sum_{j'=1}^{N}\mathrm{pr}\left\{\xi^{-1}n_1^{-1/2}\widehat{\pi}_{j,j'} > \delta N^{-1/2}\right\} \\
\leq & n_1^{-1}\xi^{-2}N\delta^{-2}\sum_{j=1}^{N}\sum_{j'=1}^{N}\mathrm{var}\left[\{Y^{\star}_{1i,j} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i},\mathbf{Z}_i^{[1]})\}\{Y^{\star}_{1i,j'} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i},\mathbf{Z}_i^{[1]})\}\right] \\
= & O(n_1^{-1}\xi^{-2}N^3),
\end{aligned}$$

which, along with Assumption (5), implies (B.10).

The proof of (B.11) is similar to that of (B.10). It is seen that

$$\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y_{1i}^{\star}(\mathbf{w}) \left\{ Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \right] \right| \\
= & \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[\sum_{j=1}^{N} w_j Y_{1i,j}^{\star} \left\{ Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \right] \right| \\
= & \sup_{\mathbf{w}\in\mathcal{W}} \left| \sum_{j=1}^{N} w_j \frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y_{1i,j}^{\star} \left\{ Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \right] \right| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \sum_{j=1}^{N} w_j \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y_{1i,j}^{\star} \left\{ Y_{1i} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \right] \right| \\
\leq & \sup_{j} \left| \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ Y_{1i,j}^{\star} Y_{1i} - Y_{1i,j}^{\star} \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\} \right| \\
= & O_p \left[\underline{n}^{-1/2} \sum_{j=1}^{N} \mathrm{var} \left\{ Y_{1i,j}^{\star} Y_{1i} - Y_{1i,j}^{\star} \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]}) \right\}^{1/2} \right] \\
= & O_p(\underline{n}^{-1/2} N)
\end{aligned}$$

by (B.1). This completes the proof. □

B.3 Proof of Theorem 4

If $\mathcal{D}^c$ is empty, the result trivially holds. We thus assume $\mathcal{D}^c$ is not empty (i.e., $M \geq 1$). For any correctly specified model $j \in \mathcal{D}$, it is obvious that

$$Y_j^{\star} - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]}) = 0. \tag{B.13}$$

We first show that

$$
\begin{aligned}
& \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1}\sum_{i=1}^{n_1}\{Y_{1i}^{\star}(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i},\mathbf{Z}_i^{[1]})\}^2 - \mathrm{E}\left[\{Y^{\star}(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}^2\right] \right| \\
= & \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1}\sum_{i=1}^{n_1}\{\sum_{j\in\mathcal{D}^c} w_j Y_{1i,j}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i},\mathbf{Z}_i^{[1]})\}^2 - \mathrm{E}\left[\{\sum_{j\in\mathcal{D}^c} w_j Y_j^{\star} - \mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}^2\right] \right|. \\
= & \sup_{\mathbf{w}\in\mathcal{W}} \left| \frac{1}{n_1}\sum_{i=1}^{n_1}[\sum_{j\in\mathcal{D}^c} w_j\{Y_{1i,j}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i},\mathbf{Z}_i^{[1]})\}]^2 - \mathrm{E}\left([\sum_{j\in\mathcal{D}^c} w_j\{Y_j^{\star} - \mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}]^2\right) \right| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \sum_{j\in\mathcal{D}^c}\sum_{j'\in\mathcal{D}^c} w_j w_{j'} \left| \frac{1}{n_1}\sum_{i=1}^{n_1}\{Y_{1i,j}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i},\mathbf{Z}_i^{[1]})\}\{Y_{1i,j'}^{\star} - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i},\mathbf{Z}_i^{[1]})\} \right. \\
& \qquad \left. -\mathrm{E}\left[\{Y_j^{\star} - \mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}\{Y_{j'}^{\star} - \mathrm{E}(Y|\mathbf{X},\mathbf{Z}^{[1]})\}\right] \right| \\
\leq & n_1^{-1/2} \sup_{j\in\mathcal{D}^c}\sup_{j'\in\mathcal{D}^c} \widehat{\pi}_{j,j'} \qquad \text{(B.14)}
\end{aligned}
$$

for any large positive constant c,

$$
\mathrm{pr}\left\{ M^{-1}\sup_{j\in\mathcal{D}^c}\sup_{j'\in\mathcal{D}^c}\widehat{\pi}_{j,j'} > c \right\} \leq c^{-2}M^{-2}\sum_{j\in\mathcal{D}^c}\sum_{j'\in\mathcal{D}^c}\mathrm{var}(\widehat{\pi}_{j,j'}) = O(c^{-2})
$$

by (B.1), and thus

$$
\sup_{j\in\mathcal{D}^c}\sup_{j'\in\mathcal{D}^c}\widehat{\pi}_{j,j'} = O_p(M). \qquad \text{(B.15)}
$$

Inserting (B.14), (B.15), (B.11) and (B.6) into (B.9), we obtain

$$
CV^{\star}(\mathbf{w}) = R^{\star}(\mathbf{w}) + O_p\{\underline{n}^{-1/2}(N^{3/2} + M)\}. \qquad \text{(B.16)}
$$

Let $\widetilde{\boldsymbol{\lambda}}$ be a weight vector with the components in $\mathcal{D}$ replaced by zeros and

$$\boldsymbol{\lambda} = \widetilde{\boldsymbol{\lambda}}/(1-\tau). \tag{B.17}$$

By (B.13) and the definition of $\boldsymbol{\lambda}$ in (B.17), we have

$$\begin{aligned} R^\star(\mathbf{w}) &= \mathrm{E}\left[\{Y^\star(\mathbf{w}) - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}^2\right] \\ &= \mathrm{E}\left(\left[\sum_{j=1}^{N} w_j\{Y_j^\star - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}\right]^2\right) \\ &= \mathrm{E}\left(\left[\sum_{j\notin\mathcal{D}} w_j\{Y_j^\star - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}\right]^2\right) \\ &= (1-\tau)^2\mathrm{E}\left(\left[\sum_{j\notin\mathcal{D}} (1-\tau)^{-1} w_j\{Y_j^\star - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}\right]^2\right) \\ &\equiv (1-\tau)^2 R^\star(\boldsymbol{\lambda}). \end{aligned} \tag{B.18}$$

Here we abused the $R^\star$ notation to denote the risk as a function of $\mathbf{w}$ and a function of $\boldsymbol{\lambda}$ as well, while the context is clear.

Combining (B.16) and (B.18), we have $CV^\star(\mathbf{w}) = (1-\tau)^2 R^\star(\boldsymbol{\lambda}) + O_p\{\underline{n}^{-1/2}(N^{3/2}+M^2)\}$ hence, replacing $\mathbf{w}$ with $\widehat{\mathbf{w}}$, we have

$$CV^\star(\widehat{\mathbf{w}}) = (1-\widehat{\tau})^2 R^\star(\widehat{\boldsymbol{\lambda}}) + O_p\{\underline{n}^{-1/2}(N^{3/2}+M)\}, \tag{B.19}$$

where $\widehat{\tau}$ and $\widehat{\boldsymbol{\lambda}}$ are τ and $\boldsymbol{\lambda}$ with $\mathbf{w}$ in their expressions replaced by $\widehat{\mathbf{w}}$. Note that in all the functions of $\mathbf{w}$ such as $R^\star(\mathbf{w})$, the expectation was calculated first and then $\widehat{\mathbf{w}}$ was inserted. Let $\widetilde{\mathbf{w}}$ be the weight vector with the first component one and other

zeros. Then, by (B.13), we know $R^\star(\widetilde{\mathbf{w}}) = 0$, which along with (B.16), implies

$$CV^\star(\widetilde{\mathbf{w}}) = O_p\{\underline{n}^{-1/2}(N^{3/2} + M)\}. \tag{B.20}$$

Now, from (B.19)-(B.20) and the truth that $\widehat{\mathbf{w}}$ minimizes $CV^\star(\mathbf{w})$, we have

$$(1 - \widehat{\tau})^2 R^\star(\widehat{\boldsymbol{\lambda}}) + O_p\{\underline{n}^{-1/2}(N^{3/2} + M)\} \leq CV^\star(\widetilde{\mathbf{w}}) = O_p\{\underline{n}^{-1/2}(N^{3/2} + M)\},$$

thus

$$(1 - \widehat{\tau})^2 \inf_{\mathbf{w}\in\mathcal{W}, \sum_{j\in\mathcal{D}} w_j = 0} R^\star(\mathbf{w}) \leq O_p\{\underline{n}^{-1/2}(N^{3/2} + M)\}, \tag{B.21}$$

which, along with Assumption 6, implies Theorem 4. □

B.4 Proof of Corollary 1

When the main model is misspecified, the results of Theorem 3 ensures that our procedure yields the minimum risk, hence is no larger than the risk of simple average to the leading order.

When the main model is correct, the simple average procedure yields

$$\widehat{Y}(\mathbf{1}/N) = \sum_{j=1}^{N} \frac{1}{N} \widehat{Y}_j = \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]}) + O_p(\underline{n}^{-1/2} N^{1/2}) + \upsilon_{N,M}, \tag{B.22}$$

where $\upsilon_{N,M}$ has the same order as MN^{-1}, so the risk is $R(\mathbf{1}/N) = O(\underline{n}^{-1}N) + \upsilon_{N,M}^2$.

On the other hand, letting $\widehat{\tau} = \sum_{j\in\mathcal{D}} \widehat{w}_j$, our prediction satisfies

$$\begin{aligned}
\widehat{Y}(\widehat{\mathbf{w}}) &= \sum_{j=1}^{N} \widehat{w}_j \widehat{Y}_j \\
&= \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]}) + \sum_{j\in\mathcal{D}} \widehat{w}_j \left\{\widehat{Y}_j - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\right\} + \sum_{j\notin\mathcal{D}} \widehat{w}_j \left\{\widehat{Y}_j - \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\right\} \\
&= \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]}) + O_p(\underline{n}^{-1/2} N^{1/2}) + O_p(1-\widehat{\tau}) \\
&= \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]}) + O_p(\underline{n}^{-1/2} N^{1/2}) \\
&\quad + O_p\left[\underline{n}^{-1/4}(M^{1/2} + N^{3/4}) \left\{\inf_{\mathbf{w}\in\mathcal{W}, \sum_{j\in\mathcal{D}} w_j = 0} R^{\star}(\mathbf{w})\right\}^{-1/2}\right] \\
&= \mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]}) + O_p(\underline{n}^{-1/2} N^{1/2}) + O_p\{\underline{n}^{-1/4}(M^{1/2} + N^{3/4})\},
\end{aligned}$$

where the second last step is based on (B.21) and the last step used Assumption 7. Hence the risk is

$$R(\widehat{\mathbf{w}}) = O_p\{\underline{n}^{-1} N + \underline{n}^{-1/2}(M + N^{3/2})\},$$

which is much smaller than $R(\mathbf{1}/N)$ as long as

$$\underline{n}^{-1} N / \upsilon_{N,M}^2 = o(1) \quad \textit{and} \quad \underline{n}^{-1/2}(M + N^{3/2})\}/\upsilon_{N,M}^2 = o(1),$$

which means $\underline{n} >> N^4 M^{-2} + N^7 M^{-4}$. □

B.5 Discussions on the variance of the averaging prediction

Let $\widehat{\mathbf{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_N)^{\mathrm{T}}$, $\boldsymbol{\Sigma} = \mathrm{diag}\{\mathrm{var}(\widehat{Y}_1), \ldots, \mathrm{var}(\widehat{Y}_N)\}$, $\boldsymbol{\Sigma}_{wy} = \mathrm{cov}(\widehat{\mathbf{w}}, \mathbf{Y})$ and $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of a matrix, respectively. We consider the following situations.

Case 1: All predictions are unbiased, i.e., $\mathrm{E}(\widehat{Y}_j) = \mathrm{E}\{\mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}$. From Proposition 3.1 of Gerda Claeskens & Wang (2016), $0 \leq \mathrm{E}(\widehat{w}_j) \leq 1$ and $\sum_{j=1}^{N} \mathrm{E}(\widehat{w}_j) = 1$, we have

$$
\begin{aligned}
& \mathrm{var}\{\widehat{Y}(\widehat{\mathbf{w}})\} \\
= \; & \mathrm{E}(\widehat{\mathbf{w}})^{\mathrm{T}} \boldsymbol{\Sigma} \mathrm{E}(\widehat{\mathbf{w}}) + 2\mathrm{E}(\widehat{\mathbf{w}})^{\mathrm{T}} \mathrm{E}\left[\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\right] \\
& + \mathrm{E}\left[\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\right]^2 - \{\mathrm{trace}(\boldsymbol{\Sigma}_{wy})\}^2 \\
\leq \; & \lambda_{\max}(\boldsymbol{\Sigma}) \|\mathrm{E}(\widehat{\mathbf{w}})\|^2 + 2\|\mathrm{E}(\widehat{\mathbf{w}})\| \left\|\mathrm{E}\left[\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\right]\right\| \\
& + \mathrm{E}\left(\lambda_{\max}[\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}^{\mathrm{T}}]\|\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\|^2\right) \\
\leq \; & \lambda_{\max}(\boldsymbol{\Sigma}) + 2\left\|\mathrm{E}\left[\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\right]\right\| \\
& + \mathrm{E}\left(\mathrm{trace}[\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}^{\mathrm{T}}]\|\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\|^2\right) \\
\leq \; & \lambda_{\max}(\boldsymbol{\Sigma}) + 2\left[\mathrm{E}\left\|\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\right\|^2\right]^{1/2} \\
& + \mathrm{E}\left\{\|\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\|^2 \|\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\|^2\right\} \\
\leq \; & \lambda_{\max}(\boldsymbol{\Sigma}) + 2\left[\mathrm{E}\left\{\|\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\|^4 \|\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\|^2\right\}\right]^{1/2} \\
& + \mathrm{E}\left\{\|\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\|^2 \|\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\|^2\right\} \\
\leq \; & \lambda_{\max}(\boldsymbol{\Sigma}) + (\boldsymbol{\Sigma}) + \left[\mathrm{E}\left\{\|\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\|^4\right\}\right]^{1/2} + 4\mathrm{E}\left\{\|\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\|^2\right\} \\
\leq \; & \max_j \mathrm{var}(\widehat{Y}_j) + 8N\left(\max_j \mathrm{E}\left[\left\{\widehat{Y}_j - \mathrm{E}(\widehat{Y}_j)\right\}^4\right]\right)^{1/2} + 4N \max_j \mathrm{var}(\widehat{Y}_j).
\end{aligned}
$$

Hence, if

$$\sqrt{n_j}[\widehat{Y}_j - \mathrm{E}\{\mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\}] \to \mathrm{Normal}(0, \sigma_j^2), \tag{B.23}$$

there exists a positive constant $\overline{c}_4$ such that

$$\sigma_j^2 \leq \overline{c}_4, \tag{B.24}$$

and

$$N\underline{n}^{-1} \to 0, \tag{B.25}$$

then the variance of $\widehat{Y}(\widehat{\mathbf{w}})$ converges to zero.

Case 2: All predictions are asymptotically unbiased with $\mathrm{E}(\widehat{Y}_j) = \mathrm{E}\{\mathrm{E}(Y|\mathbf{X}, \mathbf{Z}^{[1]})\} + O(n_j^{-1/2})$ uniformly. By basic calculations, we have

$$\begin{aligned}
& \mathrm{var}\{\widehat{Y}(\widehat{\mathbf{w}})\} \\
= \; & \mathrm{E}(\widehat{\mathbf{w}})^{\mathrm{T}}\boldsymbol{\Sigma}\mathrm{E}(\widehat{\mathbf{w}}) + 2\mathrm{E}(\widehat{\mathbf{w}})^{\mathrm{T}}\mathrm{E}\left[\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\right] \\
& + \mathrm{E}\left[\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\right]^2 - \{\mathrm{trace}(\boldsymbol{\Sigma}_{wy})\}^2 \\
& + \mathrm{E}(\widehat{\mathbf{Y}})^{\mathrm{T}}\mathrm{E}\left(\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}\left[\{\widehat{\mathbf{w}} - \mathrm{E}(\widehat{\mathbf{w}})\}^{\mathrm{T}}\{2\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\} + 2\{\widehat{\mathbf{Y}} - \mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\mathrm{E}(\widehat{\mathbf{w}})\right]\right),
\end{aligned}$$

where

$$
\begin{aligned}
& \mathrm{E}(\widehat{\mathbf{Y}})^{\mathrm{T}}\mathrm{E}\left(\{\widehat{\mathbf{w}}-\mathrm{E}(\widehat{\mathbf{w}})\}\left[\{\widehat{\mathbf{w}}-\mathrm{E}(\widehat{\mathbf{w}})\}^{\mathrm{T}}\{2\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\}+2\{\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\mathrm{E}(\widehat{\mathbf{w}})\right]\right) \\
= \;& \mathrm{E}\left(\mathrm{E}(\widehat{\mathbf{Y}})^{\mathrm{T}}\{\widehat{\mathbf{w}}-\mathrm{E}(\widehat{\mathbf{w}})\}\left[\{\widehat{\mathbf{w}}-\mathrm{E}(\widehat{\mathbf{w}})\}^{\mathrm{T}}\{2\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\}+2\{\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\}^{\mathrm{T}}\mathrm{E}(\widehat{\mathbf{w}})\right]\right) \\
\leq \;& \mathrm{E}\left[\{\sum_{j=1}^{N}O(n_j^{-1})\}^{1/2}\|\widehat{\mathbf{w}}-\mathrm{E}(\widehat{\mathbf{w}})\|\left\{\|\widehat{\mathbf{w}}-\mathrm{E}(\widehat{\mathbf{w}})\|\|2\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\|+2\|\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\|\|\mathrm{E}(\widehat{\mathbf{w}})\|\right\}\right] \\
\leq \;& (N\underline{n}^{-1})^{1/2}\mathrm{E}\{2\|2\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\|+2\|\widehat{\mathbf{Y}}-\mathrm{E}(\widehat{\mathbf{Y}})\|\} \\
= \;& O(N\underline{n}^{-1/2})
\end{aligned}
$$

under condition (4). Hence, if the conditions (4)-(B.24) are satisfied, and

$$N\underline{n}^{-1/2}\to 0, \tag{B.26}$$

then the variance of $\widehat{Y}(\widehat{\mathbf{w}})$ converges to zero.

Case 3: The main model is correct and some helper models can be misspecified. This is the case considered in Section 2.3.2. Obviously, for the correctly specified models, the predictions are asymptotically unbiased under our regularity conditions. It is seen that

$$
\begin{aligned}
& \mathrm{var}\{\widehat{Y}(\widehat{\mathbf{w}})\} \\
= \;& \mathrm{var}\left(\textstyle\sum_{j\in\mathcal{D}}\widehat{w}_j\widehat{Y}_j+\sum_{j\in\mathcal{D}^c}\widehat{w}_j\widehat{Y}_j\right) \\
\leq \;& 2\mathrm{var}\left(\textstyle\sum_{j\in\mathcal{D}}\widehat{w}_j\widehat{Y}_j\right)+2\mathrm{E}\left\{\sum_{j\in\mathcal{D}^c}\widehat{w}_j\widehat{Y}_j-\mathrm{E}\sum_{j\in\mathcal{D}^c}\widehat{w}_j\widehat{Y}_j\right\}^2.
\end{aligned}
$$

Hence, from Theorem 4, the second term above goes to zero. From the derivations in the above Case 2, the first term goes to zero. Thus, as long as the assumptions of Theorem 4 and the conditions (4)-(B.26) are satisfied, then the variance of $\widehat{Y}(\widehat{\mathbf{w}})$ converges to zero.

Case 4: The main model can be misspecified. From derivations in (B.9), (B.10), and (B.11) and Assumptions 1-3, we know that uniformly for any $\mathbf{w} \in \mathcal{W}$,

$$CV^{\star}(\mathbf{w}) = \frac{1}{n_1}\sum_{i=1}^{n_1}\{Y_{1i}^{\star}(\mathbf{w}) - \mathrm{E}(Y_{1i}|\mathbf{X}_{1i}, \mathbf{Z}_i^{[1]})\}^2 + O_p(\underline{n}^{-1/2}N^{3/2}), \tag{B.27}$$

and

$$|CV^{\star}(\mathbf{w}) - R^{\star}(\mathbf{w})| = O_p(\underline{n}^{-1/2}N^2). \tag{B.28}$$

Further, let $\mathbf{h}_j^{\star} = \left\{Y_{11,j}^{\star} - \mathrm{E}(Y_{11}|\mathbf{X}_{11}, \mathbf{Z}_1^{[1]}), \ldots, Y_{1n_1,j}^{\star} - \mathrm{E}(Y_{1n_1}|\mathbf{X}_{1n_1}, \mathbf{Z}_{n_1}^{[1]})\right\}^{\mathrm{T}}$ and $\mathbf{H}^{\star} = (\mathbf{h}_1^{\star}, \ldots, \mathbf{h}_N^{\star})$. Then,

$$CV^{\star}(\mathbf{w}) = n_1^{-1}\mathbf{w}^{\mathrm{T}}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}\mathbf{w} + O_p(\underline{n}^{-1/2}N^{3/2})$$

by (B.27). Assume that

$$\begin{gathered}\mathrm{E}\left\{CV^{\star}(\mathbf{w})\right\} \text{ has a unique minimizer } \mathbf{w}^o, i.e., \ \mathbf{w}^o = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\mathrm{E}\left\{CV^{\star}(\mathbf{w})\right\}, \\ \text{and } \mathbf{w}^o \text{ is an interior point of } \mathcal{W}.\end{gathered} \tag{B.29}$$

Let $\mathbf{u}$ be a N-dimensional vector with $\sum_{j=1}^{N} u_i = 0$ and $\|\mathbf{u}\| = 1$, and $\varepsilon_{\underline{n},N}$ be a small positive non-random value related to $\underline{n}$ and N. Then,

$$\begin{aligned}
& CV^{\star}(\mathbf{w}^o + \varepsilon_{\underline{n},N}\mathbf{u}) - CV^{\star}(\mathbf{w}^o) \\
= \; & n_1^{-1}(\mathbf{w}^o + \varepsilon_{\underline{n},N}\mathbf{u})^{\mathrm{T}}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}(\mathbf{w}^o + \varepsilon_{\underline{n},N}\mathbf{u}) - n_1^{-1}\mathbf{w}^{\mathrm{T}}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}\mathbf{w} + O_p(\underline{n}^{-1/2}N^{3/2}) \\
= \; & n_1^{-1}\varepsilon_{\underline{n},N}^2\mathbf{u}^{\mathrm{T}}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}\mathbf{u} + 2n_1^{-1}\varepsilon_{\underline{n},N}\mathbf{w}^{o\mathrm{T}}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}\mathbf{u} + O_p(\underline{n}^{-1/2}N^{3/2}) \\
\geq \; & \varepsilon_{\underline{n},N}^2\lambda_{\min}(n_1^{-1}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}) - 2\varepsilon_{\underline{n},N}(n_1^{-1/2}\|\mathbf{H}^{\star}\mathbf{w}^o\|)\lambda_{\max}^{1/2}(n_1^{-1}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}) + O_p(\underline{n}^{-1/2}N^{3/2}),
\end{aligned}$$

where

$$\begin{aligned}
n_1^{-1}\mathrm{E}\|\mathbf{H}^{\star}\mathbf{w}^o\|^2 &= \mathrm{E}\{CV^{\star}(\mathbf{w}^o) + O_p(\underline{n}^{-1/2}N^{3/2})\} \\
&= \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\mathrm{E}\left\{CV^{\star}(\mathbf{w})\right\} + O(\underline{n}^{-1/2}N^{3/2}) \\
&\leq \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\mathrm{E}\left\{R^{\star}(\mathbf{w})\right\} + \sup_{\mathbf{w}\in\mathcal{W}}\mathrm{E}\left|CV^{\star}(\mathbf{w}) - R^{\star}(\mathbf{w})\right| + O(\underline{n}^{-1/2}N^{3/2}) \\
&= \xi + O(\underline{n}^{-1/2}N^2) + O(\underline{n}^{-1/2}N^{3/2}) \\
&= O(\xi + \underline{n}^{-1/2}N^2),
\end{aligned}$$

where the last second step uses (B.28). Hence, when

$$\xi = o(1), \qquad \underline{n}^{-1/2}N^2 = o(1), \tag{B.30}$$

and when there exist positive constants $\widetilde{c}_1$ and $\widetilde{c}_2$ such that

$$\widetilde{c}_1 \leq \lambda_{\min}(n_1^{-1}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}) \leq \lambda_{\max}(n_1^{-1}\mathbf{H}^{\star\mathrm{T}}\mathbf{H}^{\star}) \leq \widetilde{c}_2, \tag{B.31}$$

we can let $\varepsilon_{\underline{n},N} = (\xi + \underline{n}^{-1/2}N^2)^{1/3}$, and obtain

$$\Pr\left\{\inf_{w^0+\varepsilon_{\underline{n},N}u\in\mathcal{W}} CV^\star(\mathbf{w}^o + \varepsilon_{\underline{n},N}\mathbf{u}) > CV^\star(\mathbf{w}^o)\right\} \to 1.$$

Now, we get if the conditions (B.29), (B.30), (B.31) are satisfied, then $\|\widehat{\mathbf{w}} - \mathbf{w}^o\| = O_p\{(\xi + \underline{n}^{-1/2}N^2)^{1/3}\} = o_p(1)$. Let w_j^o be the j^{th} component of $\mathbf{w}^o$. It is seen that

$$\begin{aligned}
\text{var}\left\{\widehat{Y}(\widehat{\mathbf{w}})\right\} &= \text{var}\left\{\sum_{j=1}^N \widehat{w}_j\widehat{Y}_j\right\} \\
&= \text{var}\left\{\sum_{j=1}^N (\widehat{w}_j - w_j^o)\widehat{Y}_j + \sum_{j=1}^N w_j^o\widehat{Y}_j\right\} \\
&\leq 2\text{var}\left\{\sum_{j=1}^N w_j^o\widehat{Y}_j\right\} + 2\text{E}\left\{\sum_{j=1}^N (\widehat{w}_j - w_j^o)\widehat{Y}_j - \text{E}\sum_{j=1}^N (\widehat{w}_j - w_j^o)\widehat{Y}_j\right\}^2.
\end{aligned}$$

Therefore, as long as the assumptions of Theorem 4 and the conditions (4)-(B.24) (B.29), (B.30), (B.31) are satisfied, then the variance of $\widehat{Y}(\widehat{\mathbf{w}})$ converges to zero.

Appendix C

Proof in Chapter 3

C.1 Proof of Theorem 5

Proof: Let $\boldsymbol{\zeta}_1 = (\boldsymbol{\tau}_1^{\mathrm{T}}, \boldsymbol{\delta}_1^{\mathrm{T}})^{\mathrm{T}}$. Let $\mathbf{B}(\cdot)$ be a vector of B-spline bases, and $\boldsymbol{\delta}_1$ satisfy $\sup_u |\boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}(u) - f_1(u)| = O(h_1^q)$, which exists under Conditions C2, C3, C4 and C5. Let

$$\boldsymbol{\phi}_i(\boldsymbol{\gamma}) = -\left[E\left\{\frac{\partial^2 \log p^I(Y, \mathbf{X}, \boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}}\right\}\right]^{-1} \frac{\partial \log p^I(Y_i, \mathbf{X}_i, \boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}.$$

$$
\begin{aligned}
\xi_i(\boldsymbol{\beta}) &= \text{expit}(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_2), \\
\eta_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma}) &= \text{expit}[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + \boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}\{p^I(0, \mathbf{X}_i, \boldsymbol{\gamma})\}], \\
\eta_{0i}(\boldsymbol{\zeta}_0, \boldsymbol{\gamma}) &= \text{expit}[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_0 + \boldsymbol{\delta}_0^{\mathrm{T}}\mathbf{B}\{p^I(1, \mathbf{X}_i, \boldsymbol{\gamma})\}], \\
\eta_{t1i}(\boldsymbol{\tau}_1, \boldsymbol{\gamma}) &= \text{expit}[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + f_1\{p^I(0, \mathbf{X}_i, \boldsymbol{\gamma})\}], \\
\eta_{t0i}(\boldsymbol{\tau}_0, \boldsymbol{\gamma}) &= \text{expit}[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_0 + f_0\{p^I(1, \mathbf{X}_i, \boldsymbol{\gamma})\}], \\
\mu_i(\boldsymbol{\beta}, \boldsymbol{\zeta}_1, \boldsymbol{\zeta}_0, \boldsymbol{\gamma}) &= \text{expit}[\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + \log\{\eta_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma})\} - \log\{\eta_{0i}(\boldsymbol{\zeta}_0, \boldsymbol{\gamma})\}], \\
\mu_{ti}(\boldsymbol{\beta}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_0, \boldsymbol{\gamma}) &= \text{expit}[\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + \log\{\eta_{t1i}(\boldsymbol{\tau}_1, \boldsymbol{\gamma})\} - \log\{\eta_{t0i}(\boldsymbol{\tau}_0, \boldsymbol{\gamma})\}], \\
\mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\zeta}_1, \boldsymbol{\zeta}_0, \boldsymbol{\gamma}) &= R_i(\mathbf{X}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}\{Y_i - \mu_i(\boldsymbol{\beta}, \boldsymbol{\zeta}_1, \boldsymbol{\zeta}_0, \boldsymbol{\gamma})\}, \\
\mathbf{U}_{ti}(\boldsymbol{\beta}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_0, \boldsymbol{\gamma}) &= R_i(\mathbf{X}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}\{Y_i - \mu_{ti}(\boldsymbol{\beta}, \boldsymbol{\tau}_1, \boldsymbol{\tau}_0, \boldsymbol{\gamma})\}, \\
\mathbf{S}_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma}) &= \{R_i - \eta_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma})\}\left[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}, \mathbf{B}\{p^I(0, \mathbf{X}_i, \boldsymbol{\gamma})\}^{\mathrm{T}}\right]^{\mathrm{T}}, \\
\mathbf{S}_{0i}(\boldsymbol{\zeta}_0, \boldsymbol{\gamma}) &= \{R_i - \eta_{0i}(\boldsymbol{\zeta}_0, \boldsymbol{\gamma})\}\left[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}, \mathbf{B}\{p^I(1, \mathbf{X}_i, \boldsymbol{\gamma})\}^{\mathrm{T}}\right]^{\mathrm{T}}, \\
\mathbf{S}_{t1i}(\boldsymbol{\tau}_1, \boldsymbol{\gamma}) &= \{R_i - \eta_{t1i}(\boldsymbol{\tau}_1, \boldsymbol{\gamma})\}\left[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}, \mathbf{B}\{p^I(0, \mathbf{X}_i, \boldsymbol{\gamma})\}^{\mathrm{T}}\right]^{\mathrm{T}}, \\
\mathbf{S}_{t0i}(\boldsymbol{\tau}_0, \boldsymbol{\gamma}) &= \{R_i - \eta_{t0i}(\boldsymbol{\tau}_0, \boldsymbol{\gamma})\}\left[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}, \mathbf{B}\{p^I(1, \mathbf{X}_i, \boldsymbol{\gamma})\}^{\mathrm{T}}\right]^{\mathrm{T}}.
\end{aligned}
$$

$$
\begin{aligned}
c_{i1} &= \mu_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma})\{1-\mu_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma})\}\{1-\eta_{t1i}(\boldsymbol{\tau}_1,\boldsymbol{\gamma})\} \\
&\quad \times[\eta_{t1i}(\boldsymbol{\tau}_1,\boldsymbol{\gamma})\xi_i(\boldsymbol{\beta})+\eta_{t0i}(\boldsymbol{\tau}_0,\boldsymbol{\gamma})\{1-\xi_i(\boldsymbol{\beta})\}], \\
c_{i0} &= \mu_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma})\{1-\mu_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma})\}\{1-\eta_{t0i}(\boldsymbol{\tau}_0,\boldsymbol{\gamma})\} \\
&\quad \times[\eta_{t1i}(\boldsymbol{\tau}_1,\boldsymbol{\gamma})\xi_i(\boldsymbol{\beta})+\eta_{t0i}(\boldsymbol{\tau}_0,\boldsymbol{\gamma})\{1-\xi_i(\boldsymbol{\beta})\}], \\
\mathbf{A}_{\boldsymbol{\zeta}_1} &\equiv \left[E\left\{I(Y_i=1)\frac{\partial \mathbf{S}_{1i}(\boldsymbol{\zeta}_1,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\zeta}_1^{\mathrm{T}}}\right\}\right]^{-1} \\
&= \left\{E\left(\xi_i(\boldsymbol{\beta})\eta_{t1i}(\boldsymbol{\zeta}_1,\boldsymbol{\gamma})\{1-\eta_{t1i}(\boldsymbol{\zeta}_1,\boldsymbol{\gamma})\}\begin{bmatrix}\mathbf{m}(\mathbf{X}_i)\\ \mathbf{B}\{p^I(0,\mathbf{X}_i,\boldsymbol{\gamma})\}\end{bmatrix}^{\otimes 2}\right)\right\}^{-1}+o(1), \\
\mathbf{A}_{\boldsymbol{\zeta}_0} &\equiv \left[E\left\{I(Y_i=0)\frac{\partial \mathbf{S}_{0i}(\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\zeta}_0^{\mathrm{T}}}\right\}\right]^{-1} \\
&= \left\{E\left([\{1-\xi_i(\boldsymbol{\beta})\}\eta_{t0i}(\boldsymbol{\zeta}_0,\boldsymbol{\gamma})\{1-\eta_{t0i}(\boldsymbol{\zeta}_0,\boldsymbol{\gamma})\}]\begin{bmatrix}\mathbf{m}(\mathbf{X}_i)\\ \mathbf{B}\{p^I(1,\mathbf{X}_i,\boldsymbol{\gamma})\}\end{bmatrix}^{\otimes 2}\right)\right\}^{-1}+o(1), \\
\mathbf{B}_{1,\boldsymbol{\zeta}_1,\boldsymbol{\gamma}} &\equiv E\left\{I(Y_i=1)\frac{\partial \mathbf{S}_{1i}(\boldsymbol{\zeta}_1,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\right\}, \\
\mathbf{B}_{0,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}} &\equiv E\left\{I(Y_i=0)\frac{\partial \mathbf{S}_{0i}(\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\right\}, \\
\mathbf{D}_{\boldsymbol{\beta}} &\equiv E\left\{\frac{\partial \mathbf{U}_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\beta}^{\mathrm{T}}}\right\} \\
&= E\left(-\mu_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma})\{1-\mu_{ti}\}[\eta_{t1i}(\boldsymbol{\tau}_1,\boldsymbol{\gamma})\xi_i(\boldsymbol{\beta})\right. \\
&\quad \left.+\eta_{t0i}(\boldsymbol{\tau}_0,\boldsymbol{\gamma})\{1-\xi_i(\boldsymbol{\beta})\}](\mathbf{X}_i^{\mathrm{T}},\mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}(\mathbf{X}_i^{\mathrm{T}},\mathbf{Z}_i^{\mathrm{T}}),
\end{aligned}
$$

(C.2)

$$
\begin{aligned}
\mathbf{D}_{\gamma} &\equiv E\left\{\frac{\partial \mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\right\},\\
\mathbf{D}_{1,\zeta_1} &\equiv E\left\{\frac{\partial \mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\zeta}_1^{\mathrm{T}}}\right\}\\
&= E\left(\mu_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma})\{1-\mu_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma})\}\{1-\eta_{t1i}(\boldsymbol{\tau}_1,\boldsymbol{\gamma})\}R_i\right.\\
&\quad \left.\times(\mathbf{X}_i^{\mathrm{T}},\mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}},\mathbf{B}\{p^I(0,\mathbf{X}_i,\boldsymbol{\gamma})\}]^{\mathrm{T}}\right)+o(1)\\
&= E\left(c_{i1}(\mathbf{X}_i^{\mathrm{T}},\mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}},\mathbf{B}\{p^I(0,\mathbf{X}_i,\boldsymbol{\gamma})\}]^{\mathrm{T}}\right)+o(1),\\
\mathbf{D}_{0,\zeta_0} &\equiv E\left\{\frac{\partial \mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial \boldsymbol{\zeta}_0^{\mathrm{T}}}\right\}\\
&= -E\left(c_{i0}(\mathbf{X}_i^{\mathrm{T}},\mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}},\mathbf{B}\{p^I(1,\mathbf{X}_i,\boldsymbol{\gamma})\}]^{\mathrm{T}}\right)+o(1).
\end{aligned}
$$

First of all, based on White (1982), the MLE $\widehat{\boldsymbol{\gamma}}$ satisfies

$$\sqrt{N}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)=N^{-1/2}\sum_{i=1}^{N}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*)+o_p(1).$$

Following the standard analysis, letting $(\widetilde{\boldsymbol{\zeta}}_1^{\mathrm{T}}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ be a point on the interval connecting $(\widehat{\boldsymbol{\zeta}}_1^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$ and $(\boldsymbol{\zeta}_1^{\mathrm{T}}, \boldsymbol{\gamma}^{*\mathrm{T}})^{\mathrm{T}}$, in step (ii), the MLE $\widehat{\boldsymbol{\zeta}}_1$ satisfies the expansion

$$
\begin{aligned}
\mathbf{0} &= N^{-1/2} \sum_{i=1}^{N} I(Y_i = 1) \mathbf{S}_{1i}(\widehat{\boldsymbol{\zeta}}_1, \widehat{\boldsymbol{\gamma}}) \\
&= N^{-1/2} \sum_{i=1}^{N} I(Y_i = 1) \mathbf{S}_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma}^*) + \left[E\left\{ I(Y_i = 1) \frac{\partial \mathbf{S}_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\zeta}_1^{\mathrm{T}}} \right\} + o_p(1) \right] N^{1/2} (\widehat{\boldsymbol{\zeta}}_1 - \boldsymbol{\zeta}_1) \\
&\quad + \left[E\left\{ I(Y_i = 1) \frac{\partial \mathbf{S}_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \right\} + o_p(1) \right] \left\{ N^{-1/2} \sum_{i=1}^{N} \boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + o_p(1) \right\}
\end{aligned}
$$

elementwise. Note that $\widehat{\boldsymbol{\zeta}}_1$ exists and is unique because of the convexity of the loglikelihood function that we maximize in Step (ii). In addition, $\|\widehat{\boldsymbol{\zeta}}_1 - \boldsymbol{\zeta}_1\| = o_p(1)$ because for any $\boldsymbol{\zeta}_1^*$ such that $\|\boldsymbol{\zeta}_1^* - \boldsymbol{\zeta}_1\| = CN^{-1/2}$, we have

$$
\begin{aligned}
& N^{-1/2}\sum_{i=1}^{N}\{R_i[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1^* + \boldsymbol{\delta}_1^{*\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\} \\
& -\log(1+\exp[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1^* + \boldsymbol{\delta}_1^{*\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}])\} \\
= \; & N^{-1/2}\sum_{i=1}^{N}\{R_i[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + \boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\} \\
& -\log(1+\exp[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + \boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}])\} \\
& +N^{-1}\{\sum_{i=1}^{N}\mathbf{S}_{1i}(\boldsymbol{\zeta}_1,\boldsymbol{\gamma})\}N^{1/2}(\boldsymbol{\zeta}_1^* - \boldsymbol{\zeta}_1) + \frac{N^{1/2}}{2N}(\boldsymbol{\zeta}_1^* - \boldsymbol{\zeta}_1)^{\mathrm{T}}\frac{\partial\sum_{i=1}^{N}\mathbf{S}_{1i}(\widetilde{\boldsymbol{\zeta}}_1^*,\widetilde{\boldsymbol{\gamma}}^*)}{\partial\boldsymbol{\zeta}_1^{\mathrm{T}}}(\boldsymbol{\zeta}_1^* - \boldsymbol{\zeta}_1) \\
\leq \; & N^{-1/2}\sum_{i=1}^{N}\{R_i[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + \boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\} \\
& -\log(1+\exp[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + \boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}])\} \\
& +\mathbf{c}^{\mathrm{T}}(\boldsymbol{\zeta}_1^* - \boldsymbol{\zeta}_1) + \frac{N^{1/2}}{2}(\boldsymbol{\zeta}_1^* - \boldsymbol{\zeta}_1)^{\mathrm{T}}\left\{E\frac{\partial\mathbf{S}_{1i}(\widetilde{\boldsymbol{\zeta}}_1^*,\widetilde{\boldsymbol{\gamma}}^*)}{\partial\boldsymbol{\zeta}_1^{\mathrm{T}}} + O_p(N^{-1/2})\right\}(\boldsymbol{\zeta}_1^* - \boldsymbol{\zeta}_1) \\
\leq \; & N^{-1/2}\sum_{i=1}^{N}\{R_i[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + \boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\} \\
& -\log(1+\exp[\mathbf{m}(\mathbf{X}_i)^{\mathrm{T}}\boldsymbol{\tau}_1 + \boldsymbol{\delta}_1^{\mathrm{T}}\mathbf{B}\{p^I(0,\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}])\},
\end{aligned}
$$

where $\mathbf{c}$ is a constant vector and we used Condition C4 in the second last step above. Note that the last term after the second last inequality above is always negative due to the convexity, so we can choose a sufficiently large C to ensure the last inequality.

Therefore, we get

$$N^{1/2}(\widehat{\boldsymbol{\zeta}}_1 - \boldsymbol{\zeta}_1) = \mathbf{A}_{\boldsymbol{\zeta}_1} N^{-1/2} \sum_{i=1}^{N} \{-\mathbf{B}_{1,\boldsymbol{\zeta}_1,\boldsymbol{\gamma}} \boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) - I(Y_i = 1)\mathbf{S}_{1i}(\boldsymbol{\zeta}_1, \boldsymbol{\gamma}^*)\} + o_p(1)$$

elementwise. Similarly we also have

$$N^{1/2}(\widehat{\boldsymbol{\zeta}}_0 - \boldsymbol{\zeta}_0) = \mathbf{A}_{\boldsymbol{\zeta}_0} N^{-1/2} \sum_{i=1}^{N} \{-\mathbf{B}_{0,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}} \boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) - I(Y_i = 0)\mathbf{S}_{0i}(\boldsymbol{\zeta}_0, \boldsymbol{\gamma}^*)\} + o_p(1)$$

elementwise, where $\boldsymbol{\zeta}_0 = (\boldsymbol{\tau}_0^{\mathrm{T}}, \boldsymbol{\delta}_0^{\mathrm{T}})^{\mathrm{T}}$.

Finally, consider the maximum pseudo-likelihood estimator $\widehat{\boldsymbol{\beta}}$ in step (iv). Obviously the loglikelihood of $\boldsymbol{\beta}$ is that of a logistic regression form which is convex, hence it has a unique maximizer $\widehat{\boldsymbol{\beta}}$ which is consistent. Then we get

$$\begin{aligned}
\mathbf{0} &= N^{-1/2}\sum_{i=1}^{N}\mathbf{U}_i(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\zeta}}_1,\widehat{\boldsymbol{\zeta}}_0,\widehat{\boldsymbol{\gamma}}) \\
&= N^{-1/2}\sum_{i=1}^{N}\mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*) + E\left\{\frac{\partial\mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\beta}}\right\}N^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
&\quad + E\left\{\frac{\partial\mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\zeta}_1}\right\}N^{1/2}(\widehat{\boldsymbol{\zeta}}_1-\boldsymbol{\zeta}_1) + E\left\{\frac{\partial\mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\zeta}_0}\right\}N^{1/2}(\widehat{\boldsymbol{\zeta}}_0-\boldsymbol{\zeta}_0) \\
&\quad + E\left\{\frac{\partial\mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_0,\boldsymbol{\gamma}^*)}{\partial\boldsymbol{\gamma}}\right\}N^{1/2}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*) + o_p(1) \\
&= N^{-1/2}\sum_{i=1}^{N}\mathbf{U}_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma}^*) + \mathbf{D}_{\boldsymbol{\beta}}N^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\
&\quad + \mathbf{D}_{1,\zeta_1}\mathbf{A}_{\zeta_1}N^{-1/2}\sum_{i=1}^{N}\{-\mathbf{B}_{1,\zeta_1,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) - I(Y_i=1)\mathbf{S}_{t1i}(\boldsymbol{\tau}_1,\boldsymbol{\gamma}^*)\} \\
&\quad + \mathbf{D}_{0,\zeta_0}\mathbf{A}_{\zeta_0}N^{-1/2}\sum_{i=1}^{N}\{-\mathbf{B}_{0,\zeta_0,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) - I(Y_i=0)\mathbf{S}_{t0i}(\boldsymbol{\tau}_0,\boldsymbol{\gamma}^*)\} \\
&\quad + \mathbf{D}_{\boldsymbol{\gamma}}N^{-1/2}\sum_{i=1}^{N}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + o_p(1).
\end{aligned}$$

Thus, $N^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \to N(\mathbf{0},\boldsymbol{\Sigma})$ in distribution, where

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathbf{D}_{\boldsymbol{\beta}}^{-1}E\left[\mathbf{U}_{ti}(\boldsymbol{\beta},\boldsymbol{\tau}_1,\boldsymbol{\tau}_0,\boldsymbol{\gamma}^*) + \mathbf{D}_{\boldsymbol{\gamma}}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*)\right. \\
&\quad -\mathbf{D}_{1,\zeta_1}\mathbf{A}_{\zeta_1}\left\{\mathbf{B}_{1,\zeta_1,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + I(Y_i=1)\mathbf{S}_{t1i}(\boldsymbol{\tau}_1,\boldsymbol{\gamma}^*)\right\} \\
&\quad \left.-\mathbf{D}_{0,\zeta_0}\mathbf{A}_{\zeta_0}\left\{\mathbf{B}_{0,\zeta_0,\gamma}\boldsymbol{\phi}_i(\boldsymbol{\gamma}^*) + I(Y_i=0)\mathbf{S}_{t0i}(\boldsymbol{\tau}_0,\boldsymbol{\gamma}^*)\right\}\right]^{\otimes 2}\mathbf{D}_{\boldsymbol{\beta}}^{-1^{\mathrm{T}}}.
\end{aligned}$$